The efficacy of problem-based learning in science education and its determinants: a study in two secondary schools in China

Siwei Xue

PhD in Science Education

Institute of Education, University College London

May 2023

Acknowledgements

This thesis is devoted to Quan Yu, my deeply cherished wife, who has comprehended and supported the most challenging husband on earth. I couldn't have achieved any of this without you.

I would like to convey my heartfelt appreciation to my supervisor, Professor Michael J. Reiss, not only for granting me the chance as a chemical engineer to undertake such educational research but also for the insightful feedback, the heartfelt motivation, and the unwavering patience. Under your guidance, I have truly learned the essence of research and intend to pursue it further. I cannot express my gratitude enough.

To my parents, Airong and Xiulian: I am grateful for the opportunities you provided by sending me from Wenzhou to Shanghai, and then to London, to obtain a superior education and broaden my perspective of the world. Thank you for the support and the sacrifices you have made. I hope my accomplishments will bring you pride.

Lastly, to myself: "This is not the conclusion. Nor is it the commencement of the end. Instead, it may well be the culmination of the start."

Declaration

I solemnly affirm that, unless specifically acknowledged, the work contained in this thesis is solely my own creation.

The word count, excluding appendices and bibliography, is 47,143.

Table of Contents

Abstract

This thesis addresses two research gaps in the field of Problem-Based Learning (PBL) by investigating four research questions through six experiments conducted in two Chinese schools over two years and nine months. With a sample size of 2,334 students from grades 8 and 9, the study demonstrates the positive impact of PBL on learning outcomes and challenges some aspects of Cognitive Load Theory (CLT). The research also examines the effects of individual PBL elements, highlighting the benefits of collective learning and PBL's synergistic effects. Moreover, it identifies factors such as prior knowledge, task complexity, previous PBL-like experiences, digital assistance, and pro-PBL family culture that can influence PBL's efficacy. These findings contribute to the field of PBL research by providing empirical evidence from a non-western cultural context, examining PBL's individual components, and identifying additional contextual factors influencing PBL efficacy. However, limitations include the generalizability of findings, measurement errors, and causality concerns. Future research should expand the scope of PBL in non-western contexts, investigate individual PBL elements and synergistic effects, measure long-term learning outcomes, and explore family-level cultural factors.

Impact Statement

This thesis has significant implications for the understanding and application of problem-based learning (PBL) in educational settings, particularly in non-western cultural contexts. By investigating the efficacy of PBL and its key elements through a series of randomized controlled experiments in Chinese secondary schools, the study provides insights into the conditions under which PBL is most effective, contributing to a more nuanced understanding of its potential benefits and challenges. The research findings also highlight the importance of considering students' prior knowledge, learning-task complexity, and other contextual factors when designing and implementing PBL, ultimately supporting the development of more effective and culturally responsive teaching practices.

By examining the synergistic effects of PBL elements, this thesis underscores the need for future research to explore the complex interactions between PBL components in order to optimize educational outcomes for diverse student populations. The investigation of individual PBL elements and their potential to improve student learning offers a more granular perspective on the relative efficacy of each component, helping to equip educators with the knowledge to tailor PBL approaches to the specific needs and contexts of their students. Furthermore, this research has the potential to foster more informed decision-making by educational policymakers, as it presents a more comprehensive understanding of the conditions under which PBL is most effective.

In addition to its implications for PBL research and practice, this thesis contributes to the broader conversation on culturally responsive education by examining the implementation and outcomes of PBL in a non-western context. As education becomes increasingly globalized, it is essential for researchers and practitioners to consider how diverse cultural contexts may shape teaching and learning experiences. By providing evidence from a Chinese context, this study not only adds to the existing body of PBL research but also encourages further examination of the ways in which educational approaches may need to be adapted and refined to better serve the needs of students in different cultural environments.

1 Introduction

Problem-based learning (PBL hereafter) is a pedagogical system that originated from the medical program of a university in North America over half a century ago (Schmidt, 2012). Since then, PBL has expanded its footprint from medical education to more domains such as science education, mathematics education, and business education (English and Kitsantas, 2019; Novak and Krajcik, 2019; Suh and Seshaiyer, 2019), and from college-based learning to curricula of broader learner bases including K-12 (Grant and Tamim, 2019).

Despite its extensive use in educational practice, the efficacy of PBL is still a controversial issue that inflames long-standing academic debates (Kalyuga and Singh, 2016; Kapur, 2016; Kirschner et al., 2018; Loibl et al., 2017; Loyens et al., 2015; Richey and Nokes-Malach, 2013; Schmidt et al., 2019; Schwartz et al., 2011; Sweller et al., 2019; Tobias et al., 2007; Tobias and Duffy, 2009). The PBL opponents like L. Zhang et al. (2022) affirm that educational policymakers around the world overstate the efficacy of PBL and ignore the prevalent empirical evidence unfavorable to the PBL approach. Even the PBL advocates such as Hung, Dolmans, et al. (2019) admit that after 50 years of research there are even more questions about PBL efficacy. Hung, Dolmans, et al. (2019) further encourage more PBL studies in non-western cultural contexts and suggest switching the research focus to why PBL is efficient or not in certain circumstances.

The present thesis echoes Hung, Dolmans, et al. (2019)'s suggestions and was undertaken to investigate the efficacy of PBL and the determinants of PBL efficacy by conducting a series of randomized controlled experiments with science learning in two secondary schools in China. To examine the PBL efficacy in the context of China, the present thesis initially formulated two research questions:

- *Research Question 1: How do PBL's key elements and its overall approach affect students' cognitive load and knowledge acquisition in learning science?*
- *Research Question 2: How do PBL's key elements and its overall approach affect students' enjoyment and self-efficacy in learning science?*

Question 1 is directly relevant to the debates between PBL advocates and opponents. The PBL advocates, sometimes using arguments based on the theory of social constructivism, hypothesize that PBL generally is beneficial to learning. However, the PBL opponents hypothesize a general

adverse impact of PBL over knowledge acquisition, sometimes using arguments drawn from cognitive load theory (CLT hereafter). CLT depicts the human cognitive architecture as a natural information processing system with limited working memory. CLT implicitly assumes that reproducing knowledge from the instructor is more efficient than inventing knowledge by the learner themself. PBL opponents argue that the guidance provided by PBL is not enough, so the learner has to employ relatively inefficient methods, such as trial-and-error or mean-ends procedures, which deplete working memory by increasing extraneous load and thus hinder the learning process (Sweller et al., 2019; Sweller, 2020).

Research Question 2 differs from Research Question 1 in terms of the aspects of learning outcomes. Previous studies evidence PBL's special effects in raising students' interest in learning and confidence in applying scientific knowledge (Areepattamannil, 2012; Cairns and Areepattamannil, 2019; Liou, 2021; McConney et al., 2014). Students' attitudes towards science can serve as a proxy for the long-term portion of learning outcome, which is not captured by short-term knowledge acquisition performance. Therefore, the present thesis examines PBL's potential long-term effects on learning performance through Research Question 2 specifically.

Following prior experiments (Kyun et al., 2013; Matlen and Klahr, 2013; Schmeck et al., 2015), the present thesis endeavours to examine the impact of PBL in a framework that incorporates both learning outcomes and cognitive load as endogenous variables. It therefore investigates not only the efficacy of the PBL approach as a whole but also the efficacy of its key individual elements. By definition, the PBL procedure is not a single teaching methodology but an amalgam of elements. The present thesis summarizes three key elements defining PBL and its variant models in practice. Based on those three key elements, three academically debatable and testable dimensions of PBL are further identified: 1) Optimal quantity of guidance, 2) Problem initiated before guidance, and 3) Small-group collaborative learning. The effectiveness of the individual pedagogical elements of PBL, as well as the PBL approach as a whole, was evaluated and compared to the traditional didactic teaching method in Experiment 1. Another noteworthy aspect of the present thesis is that it delves deeper into the examination of the synergistic effects of the PBL approach by conducting three-way variance analyses and linear regression analysis with interaction terms in addressing research questions 1 and 2. The synergistic effects of the PBL approach were assessed in Experiment 2.

Following the direction of Hung, Dolmans, et al. (2019), the present thesis does not only answer the fundamental question – "Does PBL work?" – in a non-western cultural context; it also aims to answer the more in-depth question: "When and why does PBL work?" Therefore, in addition to the first two research questions, the present thesis poses two additional research questions about the determinants of PBL efficacy:

- *Research Question 3: Do students' prior knowledge and learning-task complexity influence PBL's efficacy in terms of students' cognitive load and knowledge acquisition in learning science?*
- *Research Question 4: Do other factors influence PBL's efficacy in terms of students' cognitive load and knowledge acquisition in learning science?*

The two contextual factors in RQ3 are derived from the framework of CLT, which claims that the negative impact of PBL is more pronounced when learners have less prior knowledge or the learning task is more complex (O. Chen et al., 2017; Sweller et al., 2011). RQ4 is in line with the theoretical frameworks of modified CLTs and social constructivism which allow for more contextual factors influencing the efficacy of PBL, such as the instructor's background (Leary et al., 2013), curriculum-wide implementation of PBL (Dolmans et al., 2016), and the application of digital scaffolding techniques (Kim et al., 2018).

Experiments 3 and 4 tested the contextual effect of learners' prior knowledge and learning task complexity respectively in PBL. Experiments 5 and 6 tested other contextual factors of PBL efficacy. Particularly, the present thesis examines the contextual effects of previous PBL experience and digital scaffolding environment in Experiment 5. The present thesis further investigates the conditional effects of family-level cultural factors in Experiment 6.

The present thesis contributes to the existing literature in several ways. First, it adds new non-western-based empirical evidence to the literature on PBL. There have been limited studies (O. Chen et al., 2016; S. Gao et al., 2018) examining the effects of PBL or an inquiry-based approach (on which more below) in China, which is the country with the largest population in the world but quite different social norms from the west. The present thesis with a series of randomized controlled experiments enriches our understanding of how PBL affects the cognitive load and learning outcomes in the context of China.

Second, instead of merely treating PBL as a whole, this thesis further checks the individual effects of PBL elements. By comparing the efficacy of individual PBL elements and the PBL approach as a whole, the present thesis provides supportive evidence of the positive synergistic efficacy of the PBL approach, which is rarely mentioned by previous studies.

Third, this thesis attempts to expand upon the traditional CLT framework by exploring additional factors that may influence the efficacy of PBL. The examination of these factors, in conjunction with the traditional CLT factors of learners' prior knowledge and complexity of the learning task, will provide a more comprehensive understanding of the nature of PBL and its potential to enhance learning outcomes.

The subsequent chapters of this thesis will be organized as follows. Chapter 2 conducts an extensive review of the existing literature pertaining to studies on PBL, while also providing an overview of the educational landscape in China. Chapter 3 formulates the research questions and establishes testable hypotheses. The experimental methodology and sample formulation process are outlined in Chapter 4. The results of the experiments and the corresponding inferences are documented in Chapter 5. Chapter 6 concludes this thesis by summarizing the findings of the thesis, acknowledging its limitations, identifying its contributions, and suggesting future avenues for research.

2 Literature Review

To raise research questions and develop the hypotheses of the present thesis, this chapter reviews extant literature to figure out the theoretical and empirical gaps in current PBL-related studies. The literature review in this chapter is structured into five sections. Section 2.1 discusses how the existing literature interprets the definition of PBL. Section 2.2 explores the cognitive foundations of both PBL advocates and opponents. Section 2.3 reviews the theoretical arguments and empirical evidence about PBL efficacy with reference to the various elements of PBL. Section 2.4 enumerates the challenging issues in the empirical strategies of PBL studies. Section 2.5 discusses the institutional background of the present thesis, which is the education environment and PBL implication in China.

2.1 What is PBL?

PBL is a pedagogical system developed from various streams of educational practice in history (Darling-Hammond et al., 2020; Dolmans et al., 2016; Servant-Miklos, Woods, et al., 2019; Yew and Goh, 2016). The principle of PBL, as defined by Servant-Miklos, Norman, et al. (2019), is "the use of realistic problems as the starting point of self-directed, small-group-based learning guided by a tutor who acts as a process guide rather than a point of knowledge transfer" (p. 4). However, different versions of PBL definitions are frequently found in academic papers. For instance, PBL is "a pedagogical approach that enables students to learn while engaging actively with meaningful problems" (Yew and Goh, 2016, p. 75), and "a teaching method where the use of clinical problems is the starting point for learning, and it is through the process of working through these problems that students acquire the knowledge and skills" (Onyon, 2012, p. 22). Servant-Miklos, Norman, et al. (2019) argued that PBL is "a plastic catchall terminology" instead of a standardized and strictly defined one. The practical procedure of PBL evolved as it spread from its origin in medical school to other disciplines and from college-level students to younger learners (Jonassen, 2011). In academic communities, the definition of PBL has been a controversial topic for decades (Hung, Dolmans, et al., 2019), with the definition of PBL *per se* being seen as an 'ill-structured problem' without a standard answer.

The remaining part of Section 2.1 reviews previous studies about the PBL definition from three aspects. First, the PBL procedure is a composite of individual elements rather than a single

teaching treatment. Section 2.1.1 documents the key elements of the PBL procedure. Second, PBL is flexible and dynamic in the practice, per differing application circumstances. Section 2.1.2 scrutinizes the variance of PBL procedure standards in its practice. Third, the PBL definition has been at the crux of the debates between its advocates and opponents in the academic community. The prolonged debates shaped several determining dimensions of PBL, which are testable in the academic context. Section 2.1.3 discusses those defining dimensions from an academic perspective.

2.1.1 Key elements of PBL procedure

PBL was originally invented by educational practitioners for the medical program of McMaster University in mid-1960s Canada (Servant-Miklos, 2019b). However, PBL is now employed by pedagogical programs for various levels of learners in various disciplines, and from various cultures (Hung, Dolmans, et al., 2019), resulting in various procedures of PBL in practice. Since PBL is an amalgam of components, identifying key elements of PBL should be a precursor step to distinguishing various practical procedures of PBL.

In line with the phenomenon that there is no uniform statement for the definition of PBL, the existing documents introducing PBL provide heterogeneous versions of its distinct features. For example, Hung et al. (2008) concluded three distinct features: 1) authentic real-life clinical problems, 2) students actively engaging in self-directed problem-solving, and 3) learning processes in small-group settings (p. 496). Savery (2015) suggested three defining elements of PBL: 1) the tutor as a facilitator of learning, 2) the learners to be self-directed in their learning, and 3) the essential elements in the design of ill-structured instructional problems (p. 15). Schmidt et al. (2019) rather advised six defining elements of PBL: 1) the problems as the starting point for learning, 2) students collaborating in small groups, 3) flexible guidance of a tutor, 4) a limited number of lectures, 5) student-initiated learning, and 6) ample time for self-study (p. 32). Savery (2019) even listed 16 characteristics to distinguish PBL and other learning paradigms. Schmidt (2012) and Servant-Miklos, Norman, et al. (2019) reviewed the history of PBL development and figured out the most influential educational and psychological roots as follows:

1) John Dewey's experiential learning inspired the use of 'Problem';
2) Carl Rogers's Humanist psychology encouraged 'Self-directed learning'; and

3) Constructivist psychology motivated ‘Exploratory group discussion’ and ‘Scaffolding by the tutor.’

The above-mentioned framework provided by Servant-Miklos, Norman, et al. (2019), incorporating other versions of distinct features from Hung et al. (2008), Schmidt (2012), and Savery (2015), is consistent with the three distinct features defined by Hung, Dolmans, et al. (2019): (1) problem-initiated and problem-driven learning, (2) self-directed learning with tutor facilitation, and (3) collaborative learning in small groups (p. 945). The present thesis defines PBL as a practical pedagogical methodology with the three key elements defined by Hung, Dolmans, et al. (2019).

2.1.2 The various practical procedures of PBL

2.1.2.1 Two camps of PBL models in medical education

The majority of current research on PBL efficacy remains in the medical field (Moallem, 2019). PBL proponents in the medical discipline (e.g., Servant-Miklos, 2019a) partitioned the practical PBL procedures into two camps: McMaster curricula and Maastricht curricula,[1] which are denoted as Type 2 and Type 1 PBL curricula respectively (Schmidt et al., 2009; Schmidt, 2012). According to Servant-Miklos (2019c), both types of PBL models are still prevalent in college programs.

The Medical program of McMaster University in Canada, which was launched in the mid-1960s, was the first curriculum to adopt the PBL conception (Servant-Miklos, 2019b). Servant-Miklos, Norman, et al. (2019) argued that after 1977, McMaster PBL curricula were influenced by Howard Barrows and deviated from their original principles. Servant-Miklos, Norman, et al. (2019) denoted McMaster PBL curricula as Type 2 PBL curricula. Type 1 PBL curricula, the Maastricht model, originated at the Maastricht University in 1970s (Servant-Miklos, 2019a). Unlike the McMaster program, the Maastricht program initially targeted inexperienced students in medical school, which steered it to be a more standardized paradigm in terms of problem design and evaluation.

[1] Strictly speaking, there exists the third camp of PBL originating from Aalborg University. However, it deviates more from the core principles of PBL and is less representative. Servant-Miklos and Spliid (2017) provide more details.

In addition to a higher level of standardization, Maastricht's PBL model emphasizes more the tutor's scaffolding function. Conversely, the McMaster PBL model offers more freedom to students' self-directed learning (Neville et al., 2019). The problem designed by Maastricht's PBL model is more likely to lead students to the intended learning issue. Furthermore, the Maastricht PBL model promotes small-group discussion more than the McMaster PBL model (Schmidt and Mamede, 2020). Table 2.1 compares the components of the McMaster and Maastricht models.

Table 2.1: Comparison between McMaster and Maastricht models

PBL Element	Maastricht model (Type 1 PBL)	McMaster model (Type 2 PBL)
Problem-initiated and problem-driven learning	More likely to lead students to the intended learning issue	Less likely to lead students to the intended learning issue
Self-directed learning with tutor facilitation	Students have less freedom in self-directed learning; Emphasis on tutor's scaffolding	Students have more freedom in self-directed learning; Less emphasis on tutor's scaffolding
Collaborative learning in small groups	More encouraged	Less encouraged

Source: Compiled by author

A strand of research (Schmidt et al., 2009; e.g., Schmidt, 2012; Servant-Miklos, 2019c, 2019b) has investigated the theoretical divergence between the McMaster and Maastricht PBL models. Servant-Miklos (2019c) argued that the goals of the McMaster and Maastricht PBL models are contrary. McMaster's PBL model tends to develop students' problem-solving skills, and assumes that the skill of problem solving is independent of the content. The Maastricht PBL model inclines toward facilitating students' knowledge acquisition. It is however not the focus of the present thesis to determine which of the two PBL models is more authentic. The relevant implication here is that even in the field of medical education, where PBL originated, there is considerable room for alterations in the practical procedure.

2.1.2.2 More PBL models beyond the domain of medical education

As the footprint of PBL expanded from medical education to more domains such as science, mathematics, and business (English and Kitsantas, 2019; Novak and Krajcik, 2019; Suh and Seshaiyer, 2019), and from college-based learning to curricula of broader learner bases including K-12 cultivation (Grant and Tamim, 2019), the PBL models varied to a larger extent.

In both the McMaster and Maastricht models, the initiated problem should be ill-structured and call for decision-making, since the goals of the two curricula are to develop learners' professional skills in applying knowledge. In other non-medical disciplines, the initiated problem of PBL can be chosen from a typology of problems on a continuum from well-structured to ill-structured (Jonassen and Hung, 2015). For example, the PBL for mathematics and the sciences is more likely to use well-structured story problems (Jonassen, 2011, p. 96), departing from both McMaster or Maastricht models.

The usages of guidance in both McMaster and Maastricht curricula are also restricted, though the latter emphasizes more a tutor's scaffolding, as presented in Table 2.1, because the initially targeted learners of the two programs are college students. But for a broader application environment, Jonassen (2011) argued that "learners' levels of prior knowledge, experience, reasoning ability, various cognitive styles, and epistemic beliefs" affect problem-solving (p. 96). Therefore, PBL can choose the guidance methods from a bundle of scaffolding tools including high-level guidance methods such as worked examples and structural analogs (p. 101). Moallem (2019) claimed that "the contextual factors (e.g., setting, content, age group, time) effect changes in the design of the PBL process and the degree of the facilitator's support during various phases of PBL" (p. 120), reinforcing that there is no single PBL model fit for all the learning contexts. When targeting inexperienced learners such as K-12 students, PBL's scaffolding should provide even more guidance.

One consequence of the increasingly variant PBL model is that its boundary with other learner-centered learning methodologies becomes vague. Savery (2019), Moallem (2019), and Wijnia et al. (2019) tended to differentiate PBL from Project-based learning, Discovery learning, Case-based learning, Learning by design, and Inquiry-based learning. However, the identified differences are too trivial to force those learner-centered learning approaches to violate the three key principles of the PBL procedure listed in Section 2.1.1.[2] Those inquiry-based learning approaches thus can be broadly interpreted as varying PBL models adaptive to differing

[2] For example, Savery (2019) advised that the problems in the Project-based learning method are usually well-structured (closed vs. open-ended) with varying degrees (p. 100). Moallem (2019) presented that problems of Inquiry-based learning are the driving questions created by the teacher (p. 118). See more comparing details in Table 4.7 of Savery (2019), Table 5.2 of Moallem (2019), and Table 12.2 of Wijnia et al. (2019).

curricula. As Moallem (2019) (p. 110) said, for PBL, "at one extreme a problem might be ill-structured where context is crucial, solutions may not even exist, and evaluation is more about the evidence and chain of reasoning employed than the solution itself. At the other extreme, a problem might be highly structured with a focus on accurate and efficient paths to an optimal solution where context is a secondary concern." Such interpretation is shared by PBL opponents who view PBL and other inquiry-based learning approaches as the same methodology with different names (Kirschner et al., 2006; Sweller et al., 2019). The present thesis also concurs with such interpretation and perceives those learner-center learning methods essentially as variants of PBL.

2.1.3 The academic perspective of defining PBL

In a review of the meta-analyses of the past 50 years of research on PBL, Hung, Dolmans, et al. (2019) concluded that "Fifty years of research has given us a better understanding of PBL" (p. 952). Hung, Dolmans, et al. (2019) elucidated the fifty years of research on PBL as three waves of debates between PBL advocates and opponents in the academic community. Those enduring debates[3] identified several testable features of PBL and shaped the academic perspective of defining PBL.

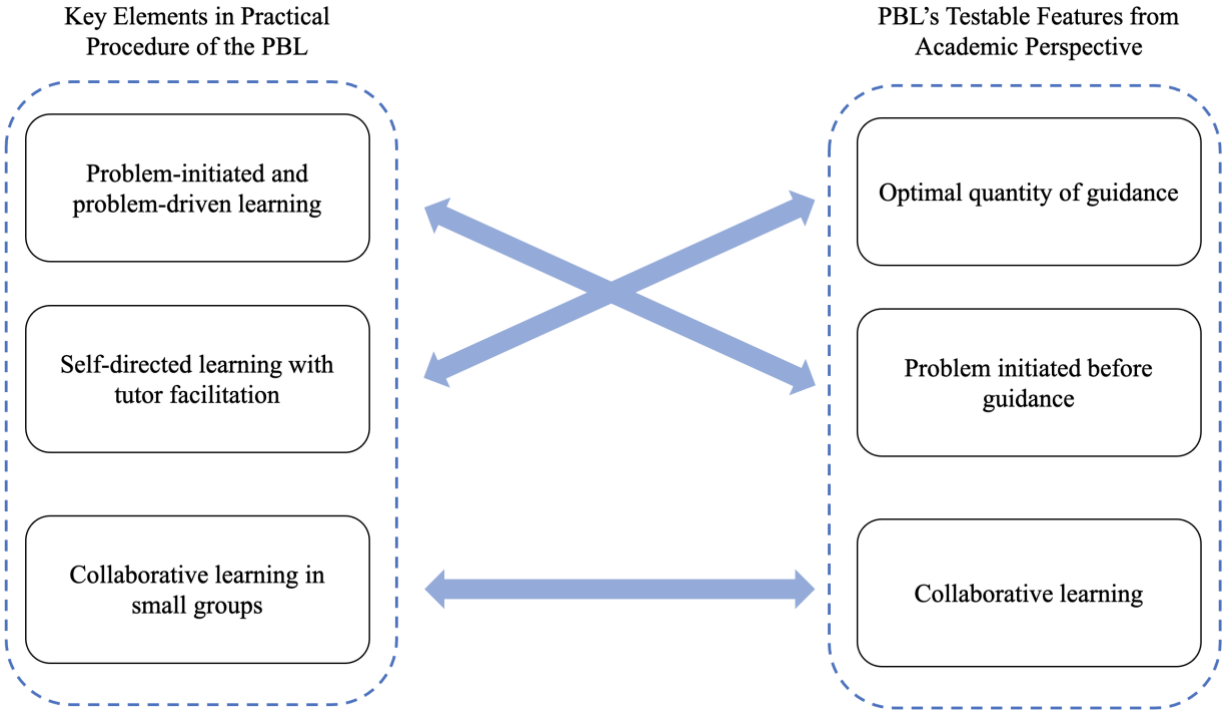
Partially inspired by the framework suggested by Wise and O'Neill (2009) which divided the debating issues into the quantity, the timing, and the context of guidance, together with other review studies including L. Zhang (2016) and Hung, Dolmans, et al. (2019), the present thesis summarizes the following three debatable and testable features of PBL from an academic perspective:

1) Optimal quantity of guidance,
2) Problem initiated before guidance, and
3) Small-group collaborative learning.

[3] See Tobias et al. (2007), Tobias and Duffy (2009), Schwartz et al. (2011), Richey and Nokes-Malach (2013), Loyens et al. (2015), Kalyuga and Singh (2016), Kapur (2016), Loibl et al. (2017), Kirschner et al. (2018), Schmidt et al. (2019), and Sweller et al. (2019) for an overview of those ongoing debates.

The above three testable features correspond to the three key elements of PBL procedures in Section 2.1.1. Figure 2.1 shows the correspondent relation between the key elements of practical procedures and testable features from the academic perspective.

Figure 2.1: The interaction between the practical and academic perspectives



Source: Compiled by author

As Figure 2.1 shows, a one-to-one correspondence can be built between the key elements of PBL's practical procedures and the academically testable features of PBL. The element of problem-initiated and problem-driven learning determines the timing of guidance, which should occur after the problem. The element of self-directed learning with tutor facilitation or scaffolding means the quantity of guidance should not be at the maximum level, as in the traditional didactic teaching method. The element of collaborative learning is literally the same between the practical and academic perspectives.

2.1.3.1 Optimal quantity of guidance

The first academically testable feature of PBL is based on the dimension of guidance quantity, which may be the most controversial part of the PBL definition in academic debates. The second wave of research, as identified by Hung, Dolmans, et al. (2019), was instigated by Kirschner et

al. (2006)'s challenge on the quantity of guidance in PBL. PBL opponents such as Kirschner et al. (2006) and Sweller et al. (2007) argued that PBL provides no or minimal guidance to the learner. In contrast, Hmelo-Silver et al. (2007) and Schmidt et al. (2007) claimed that the tutor in PBL does provide extensive and meaningful instruction to the learner. Thus the discussion became less productive because the two sides hold inconsistent definitions of PBL. To avoid unsolvable disputes in the definition related to PBL guidance quantity, a number of subsequent studies created new terminologies, such as unassisted discovery (Alfieri et al., 2011), or used terminology previously employed in other contexts, such as guided inquiry (Roll et al., 2018).

The instruction provided by PBL is the scaffolding element mentioned in Section 2.1.1. PBL proponents such as Schmidt et al. (2007) believe that scaffolding in PBL allows for flexible adaption of guidance so that guidance in PBL is not minimal but optimal. Schmidt (2012) suggested that PBL instructors should optimally choose the guidance quantity within a wide spectrum, as long as the solution is not explicitly revealed. The optimal quantity of guidance provided by PBL is also named 'just-right-amount' by PBL advocates, conditioning on the merits of knowledge and student taught.

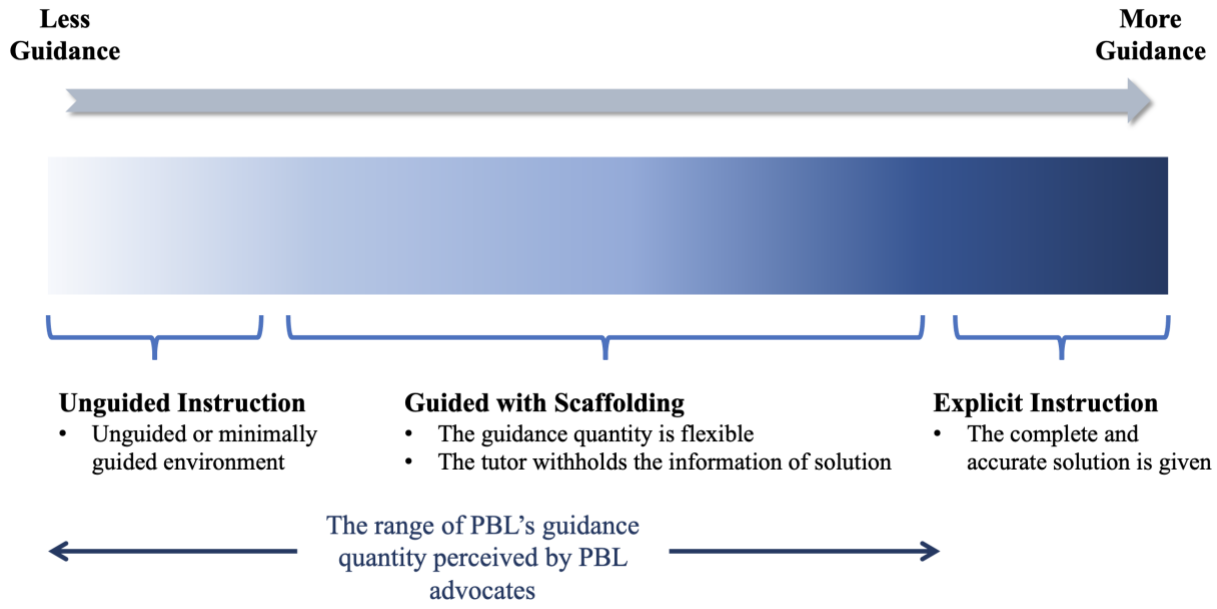
On the other hand, the opponents of PBL such as Sweller (2009) and Clark (2009) insisted that PBL has a tendency not to offer students 'accurate and complete' information. They further assume that PBL is a teaching method with minimal guidance. Such controversial understanding has been further exaggerated by subsequent studies to the point that it is sometimes maintained as long as the lecture's instruction exists, it is neither PBL nor an inquiry-based learning approaches.[4]

Despite the inconsistent understanding, a consensus between PBL advocates and opponents has gradually been established (Kelly, 2014; Matthews, 2015; Sweller, 2020; Xu et al., 2021; Zambrano et al., 2019b) that as long as the scaffolding does not explicitly offer complete information or provide maximum-level guidance, it still belongs to the PBL territory. Therefore,

[4] For instance, PISA2015-based studies such as Jerrim et al. (2020) and Liou (2021) usually exclude two items in PISA's IBTEACH variable in their analyses since the two items are about the teacher's instruction. See more details in the discussion of Section 2.4.2.

the spectrum of guidance quantity from PBL to explicit instruction could be depicted as in Figure 2.2.

Figure 2.2: A spectrum of guidance quantity from PBL to explicit instruction



Source: Compiled by author

Figure 2.2 illustrates how PBL occupies a wide interval in the guidance quantity spectrum compared to the traditional didactic teaching scheme. That interval includes guidance methods such as comments, feedback, suggestions, and explanations in the PBL process (Ertmer and Glazewski, 2015, 2019). PBL aims to achieve the optimal degree of guidance within this interval. It is worth noting that a worked example is also an available guidance option for PBL, as clearly stated by Jonassen (2011). However, PBL opponents consistently ignore that and regard the worked-example guidance as an indication of explicit instruction (Kalyuga and Singh, 2016; Kyun et al., 2013; L. Zhang and Cobern, 2021). The present thesis agrees with PBL advocates' definition that PBL tends to choose the optimal level of guidance quantity within a wide range of possibilities including the use of worked examples.

2.1.3.2 Problem initiated before guidance

The second academically testable feature of PBL is based on the dimension of guidance timing, or the sequence of problem and guidance. In most definition statements of PBL (Darling-Hammond et al., 2020; Servant-Miklos, Woods, et al., 2019; Yew and Goh, 2016), the problem

should be introduced first and be the starting point of a learning cycle. When instruction is offered before the learner's problem-solving attempt, even if the information is withheld in instruction, it is still not PBL. Tawfik et al. (2020) specifically stated that the pedagogical method that has a lecture about the problem prior to problem-solving is not PBL. Thus, having the problem initiated before guidance becomes another determining feature of PBL.

The sequence of guidance is a dichotomy variable depending on whether the problem comes first or not. So, it is more easily amendable to the empirical and experimental testing designs, compared to the optimal quantity of guidance which is more subject to individual judgment. There has been a variety of terminologies describing the guidance sequence. Hsu et al. (2015) differentiated WP (worked-example–problem) and PW (problem–worked-example sequence) sequences. Chase and Klahr (2017) contrasted IT (invent then tell) and TP (tell then practice) instructional methods. Loibl et al. (2017) and Sinha and Kapur (2021) identified PS-I (problem solving followed by instruction) and I-PS (instruction followed by problem solving) designs. Whatever specific terminology is used, the advocates of PBL (Chase and Klahr, 2017; Kapur, 2016; Sinha and Kapur, 2021) and the opponents of PBL (Ashman et al., 2020; Hsu et al., 2015; Matlen and Klahr, 2013) have little disagreement on this academically defining feature.

A potential disputing issue about guidance timing is how to define the starting point of a learning cycle. For instance, Klahr and Nigam (2004) conducted an experiment and found a group of students with direct instruction learn the principle better. But before the investigated learning cycle, there was a phase for baseline assessment. If we alternatively treat the assessment phase as the starting point of the learning cycle and interpret baseline assessment as raising a problem for students, the lecture-problem sequence in Klahr and Nigam (2004)'s experiment could be reverted. That is why for the same experiment results of Klahr and Nigam (2004), Kapur (2016) inferred the opposite conclusions that "the very effects that Klahr and Nigam attribute to direct instruction alone seem more appropriately attributed to a pure discovery learning phase (their baseline assessment) followed by direct instruction" (p. 291). More detailed discussions about this issue are in Section 2.4.3 of the present thesis.

2.1.3.3 Small-group collaborative learning

The last academically testable feature of PBL is based on the dimension of collaborative learning. The PBL procedure requires small-group discussion but does not specify the exact

group size. The group size of PBL could be 4 to 6 students in the traditional McMaster program (Servant-Miklos, Norman, et al., 2019) or as many as 10 in certain situations (Schmidt et al., 2019, p. 27). It could be roughly assumed that it deviated from the spirit of PBL if the group size exceeds 10. Some experimental studies (Zambrano et al., 2019b, 2019a) set the size as 3.

Besides group size, previous literature figured out that other factors, like group member diversity or group climate, affect the efficacy of collaborative learning (Fonteijn and Dolmans, 2019). However, other factors are not as emphasized by the PBL procedure as is group size. On the PBL element of collaborative thinking, there is no prominent difference between the supporters and opponents of PBL. They even have the same positive prediction of the effect of collaborative thinking. The differences stem from their underlying theoretical explanations. This issue will be further discussed in Section 2.3.3. The present thesis defines the last academically testable feature of PBL as collaborative learning with each group smaller than 10 persons.

2.1.4 Summary of Section 2.1

The definition of PBL is not as black and white as a factual statement. The inconsistency of interpreting PBL by scholars breeds a significant portion of disputes over PBL efficacy. Therefore, the literature review of the present thesis uses one separate section to review and discuss how to define PBL from both practical and academic perspectives.

In the practical procedure, PBL is not a single teaching methodology but an amalgam of elements. Particularly, three key elements of the PBL procedure could be identified: (1) problem-initiated and problem-driven learning, (2) self-directed learning with tutor facilitation, and (3) collaborative learning in small groups. Following those three key elemental principles, PBL evolved into a plethora of variant models from its original McMaster and Maastricht forms. Those PBL models vary with a typology of problems and different levels of guidance to meet the needs of new curricula beyond the medical education domain. Those variants of PBL models include other learner-center learning methods, like Discovery learning, Project-based learning, and Inquiry-based learning, which despite superficial incongruity share the core principles of PBL.

Academic debates in recent decades have been conducive to our better understanding of PBL. From an academic perspective, several debatable and testable defining features of PBL were

shaped. This section summarizes the three academically testable features of PBL: 1) Optimal quantity of guidance, 2) Problem initiated before guidance, and 3) Small-group collaborative learning. Those three academically debatable and testable features, which correspond to the three key elemental principles of the PBL procedure, constitute the framework for reviewing the theoretical arguments and empirical evidence of PBL efficacy in Section 2.3. Breaking down PBL into debatable and testable features is also essential to mitigate the pitfalls of experimental approaches discussed in Section 2.4 and eventually shapes the methodological design of the present thesis in Chapter 4.

2.2 Cognitive foundation of PBL efficacy

One interesting point in the debates over PBL efficacy is that the two sides build their theoretical arguments on different cognitive foundations. The cognitive foundation involves presumptions about human cognitive architecture and the process of human cognitive development. Different cognitive foundations result in different systems of terminologies when describing the same thing and generate different theoretical hypotheses. Therefore, the cognitive foundation is a *sine qua non* of the theoretical arguments of PBL efficacy and will be explored in the current section.

2.2.1 Traditional cognitive load theory (CLT)

Traditional CLT describes learning as a process of acquiring new knowledge and portrays the human cognitive architecture as a natural information processing system (Sweller et al., 2011, 2019). The traditional CLT defines knowledge by borrowing insights from evolutionary psychology, which classifies knowledge into two genres, biologically primary and biologically secondary knowledge (Geary, 2008, 2012; Geary and Berch, 2016). The difference is that primary knowledge does not need to be learned, while secondary knowledge needs to be learned.

The reason humans does not need to learn biologically primary knowledge is that the survival rule already disciplined us to obtain that knowledge through evolution. Traditional CLT even argues that general problem-solving skills belong to biologically primary knowledge and thus people don’t need to learn them (Sweller, 2021, p. 2). On the other hand, biologically secondary knowledge has to be learnt after birth. Under the CLT framework, there are two ways of learning knowledge. When the knowledge can not be copied from others, we have to ‘invent’ it by trial-and-error method or means-ends analysis. Otherwise, we can directly transfer knowledge from

others. A critical proposition of CLT is the latter way of learning knowledge (i.e., transferring from others) is always more efficient (Sweller, 2021).

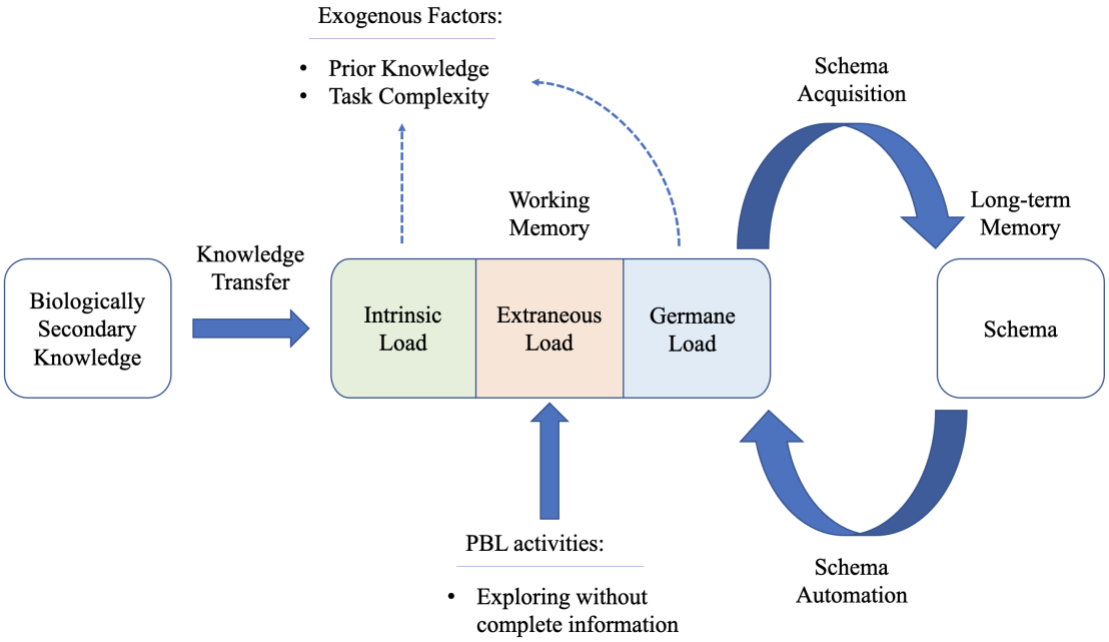
Transferring knowledge is a process that involves both the learner's prior and new knowledge. The prior knowledge is stored in long-term memory or domain-specific schemas which can be activated effortlessly, i.e., at no cost (Sweller, 2020), whereas the new knowledge needs to be processed by working memory. Compared to long-term memory, working memory has limits in terms of both duration and capability (Martin, 2016). When working memory is cognitively overloaded, the human information processing system stops transferring knowledge to long-term memory (Sweller et al., 2011, 2019).

So under the traditional CLT framework, the efficiency of learning is determined by the cognitive load, given a fixed working memory limit. The more cognitive load, the less efficient the learning. CLT further claims that both relevant and irrelevant activities would increase the cognitive load, which are denoted as intrinsic and extraneous loads respectively. Traditional CLT scholars further suggested that the resources of working memory for acquiring and automating schemas are different from either intrinsic or extraneous loads. They denoted it as germane load.[5] As extraneous load represents the waste of limited working memory resources, the standard of assessing a learning methodological design by CLT-based studies is whether it increases or reduces extraneous loads (Martin, 2016; Martin and Evans, 2018, 2019; Sweller, 2009).

To better interpret the conceptual framework of traditional CLT, Figure 2.3 indicates that a variety of learning activities occur simultaneously to compete for limited working memory. Knowledge transfer consumes intrinsic load. Schema automation and acquisition consume germane load. Both intrinsic and germane loads are necessary for achieving learning goals and are determined by the learner's prior knowledge and the learning task's complexity. In the view of traditional CLT, PBL activities such as letting the learner explore with incomplete information lead to extraneous load and adversely affect learning efficiency.

[5] However, not all CLT scholars agree on this three-load classification. For instance, Kalyuga (2011) suggested that germane load refers to the same consumed resources with the intrinsic load. Using rating experiments, D. Jiang and Kalyuga (2020) presented evidence that intrinsic and germane loads are highly correlated, supporting the two-load classification argument.

Figure 2.3: The conceptual framework of traditional CLT



Source: Compiled by author

2.2.2 Social constructivist learning theory

In contrast to CLT, one branch of PBL advocates employing the social theory of constructivist learning or constructivism to explain learning (Hung, Moallem, et al., 2019). First, constructivism theory understands the learning process as acquiring (making) new knowledge from the learner's inside instead of transferring knowledge from the learner's outside (Schmidt et al., 2019). The term 'constructivist' therefore stems from such a learner-centered view of the learning process. Second, constructivism emphasizes the role of social interaction in learning. This social interaction facilitates each learner's learning motivation (Russo and Hopkins, 2019), awareness of knowledge gap(s) (Newman and DeCaro, 2018), conceptual change (Loyens et al., 2015), and deep knowledge construction (Allen et al., 2013; Chin et al., 2016).

Social constructivist learning theory implicitly assumes several principles of the cognitive foundation. The first principle is that the individual mutation of knowledge in learning is encouraged. Under CLT which I discussed in Sector 2.2.1, the purpose of learning is to 'copy' knowledge. Although CLT also admits the existence of knowledge reorganization, it regards the mutation of knowledge as it is transferred as the noise of learning (Sweller et al., 2019).

Constructivism scholars, however, view the world as "multiple, changing realities where the knower cannot be separated from the known" (Hung, Moallem, et al., 2019, p. 52). Therefore, social constructivist learning theory encourages learners to create their own version of knowledge.

The second principle of constructivism is that, unlike CLT where the knowledge is context-independent, certain knowledge must be demonstrated within its context. As Hung, Moallem, et al. (2019) put it, "learning is not only situated within the immediate environment surrounding the individual but within a much broader social-cultural environment" (p. 56). The knowledge learned from the same classroom will vary because of the learner's individual past experience, cognitive puzzlement, and social negotiation. Because the social environment is part of knowledge, one advantage of PBL is that it can mimic the social environment where the students are going to apply the knowledge. Thus, the type of guidance such as comments, feedback, suggestions, and explanations matter in the PBL process (Ertmer and Glazewski, 2015, 2019). Such an argument is also named the 'situated learning hypothesis.'

The third principle of constructivism is that for certain 'deep' knowledge it is more efficient to be absorbed through the learner's inventing instead of instruction's lecturing. The instructor can facilitate the learner's invention of knowledge. Schwartz et al. (2011) provided evidence that a contrasting case is beneficial for the learner's inventing knowledge. Chin et al. (2016) suggested that the inventing task assigned by the instructor is more efficient than the traditional 'compare and contrast' approach in terms of the learner's resulting learning. Such a principle of constructivism is also against traditional CLT. Traditional CLT says that the trial-and-error method or means-ends analysis invents the knowledge with an element of randomness, which is always less effective. The emerging modified CLT tries to integrate the discrepancy between these two approaches, as I discuss in Section 2.2.4.

A significant drawback of social constructivist learning theory is that its model is not standardized and explicit. Compared to CLT which specifies the mechanism of cognitive load, it is difficult to empirically test social constructivist learning theory. The theoretical origins of Constructivism, according to Hung, Moallem, et al. (2019), include Piaget's Cognitive Equilibrium Theory, Vygotsky's Sociocultural Constructivism, Activity Theory, and Situated Learning. Those scattered theoretical origins do not integrate into a model clearly describing the

learner's cognitive development. Tobias and Duffy (2009) even claimed that "Constructivism remains more of a philosophical framework than a theory that either allows us to precisely describe instruction of prescribing design strategies" (p. 4). Many empirical studies supporting the efficacy of PBL do not explicitly mention constructivism (Darabi et al., 2018; Newman and DeCaro, 2018; Russo and Hopkins, 2019). Rather, they focus on a specific issue such as motivation or awareness of the knowledge gap.

2.2.3 The challenges to traditional cognitive load theory

Compared to social constructivist learning theory, the strength of the CLT cognitive foundation is that it is a concrete model. However, CLT's cognitive model might be oversimplified. CLT describes learning as a process with the single purpose of changing long-term memory. Once the knowledge is stored in the learner's long-term memory, the learning process is done. The only issue inhibiting knowledge absorption is the cognitive load due to the working memory limit. In traditional CLT, the cognitive load is, however, only affected by two factors: the learner's prior knowledge and the complexity of the learning task (Brünken et al., 2010; Kalyuga and Singh, 2016).

A number of educational practitioners believes that learning is a more complicated process than is described by CLT. Jonassen (2009) (p. 13) claimed that CLT "focuses only on working memory and long-term memory, ignoring all other cognitive constructs. A cognitive architecture must account for the context, the learner, and the processes of cognition (social and cognitive) in order to explain or predict cognitive activities." Taber (2013) reviewed the theoretical models of thinking, understanding, and learning. His review suggested that a learning-as-information-transfer model as in CLT is too simplistic. "Knowledge is not just information that can simply be transmitted as long as transmitter and receiver are functioning well and clear lines of communication have been established" (p. 278). This section documents several challenging issues to traditional cognitive load theory.

2.2.3.1 The missing influential factors in the CLT model

In the CLT model, if extraneous load remains constant by fixing the teaching methodology, learning efficacy is determined by only two factors: the learner's prior knowledge and the learning task's complexity. However, evidence from empirical studies, especially association analyses using large samples, suggests that learner's prior knowledge and learning task's

complexity can only explain a small portion of the variance in learning outcomes. For example, Cairns and Areepattamannil (2019) conducted a large-sample-based cross-country study, which showed that country-level factors explained 21% of the total variance in learning outcomes higher than the 18% of student-level factors. Areepattamannil (2012) showed that within-school factors together explained less than 10% of the variance in learning outcomes. In the regressional analysis of S. Gao et al. (2018), which controlled for factors associated with CLT, the adjusted R^2 was only 0.034. Several influential factors affecting learning outcomes seem to be missed in the traditional CLT model.

From Figure 2.3, we can induce two layers of missing factors: the missing factors which influence learning outcomes through cognitive load and the missing factors which directly influence learning outcomes. Section 2.2.3.1 mainly focuses on the first layer of missing factors while Section 2.2.3.2 focuses on the second layer. The missing factors of cognitive load, as suggested by existing literature, could include motivation, emotion, awareness of the knowledge gap, and confidence about prior knowledge.

Motivation is one of the most prominent factors ignored by the traditional CLT model. Traditional CLT believers treat motivation as a process of preparing mental resources which proceed to the learning phase (Feldon et al., 2019). CLT scholars either assume that once moving into the learning phase the learner already prepares necessary mental resources or assume that the learner's primary knowledge can affect the motivation through the expected difficulty of learning (Lespiau and Tricot, 2018, 2019). Therefore motivation becomes a mediating variable that can be pruned away from the CLT model.

The story would be different in two scenarios where 1) motivation is determined by factors beyond the learner's prior knowledge or learning task complexity, and 2) motivation can directly influence the learning outcome instead of through mental resources or working memory limits. The current section concentrates on the former scenario. First, evidence provided by Rey and Steib (2013), Kalyuga and Singh (2016), and Plass and Kalyuga (2019) suggests that emotion influences motivation. Furthermore, social constructivists argue that challenging problems (Loibl et al., 2017; Russo and Hopkins, 2019) or productive failures (Darabi et al., 2018; Wijnen et al., 2018) can boost learners' motivations. In summary, motivation and its determinants beyond cognitive load could be another missing factor in the traditional CLT model.

Emotion is also not included in the traditional CLT model (Brünken et al., 2010). There are studies (Baddeley, 2012; Fraser et al., 2015) indicating that negative emotion can reduce the overall limit of working memory and thus the cognitive load. A different interpretation of such evidence is that emotion causes extraneous load while the working memory limit remains the same. Consistent with the assumption that emotion contributes to extraneous load, some scholars argue that even positive emotion adversely influences learning (Park et al., 2011; Pekrun and Linnenbrink-Garcia, 2012). Conversely, research from the field of cognitive architecture, such as Zlomuzica et al. (2016), suggests that positive emotion is associated with better spatial memory. There are also scholars, such as Plass and Kalyuga (2019), who advise that emotion *per se* does not consume working memory. The impacts of emotion on learning are through motivation which was discussed in the paragraph above on motivation. Based on whichever above-mentioned argument, emotion is at least a factor influencing working memory, in addition to prior knowledge and task complexity.

Awareness of the knowledge gap is closely related to learners' motivations. Social constructivists argue that humans have an incentive to achieve equilibrium between their inside and outside knowledge set (Schmidt et al., 2019). Therefore, awareness of the knowledge gap brings learners motivation. In traditional CLT, recognizing the knowledge gap seems to be automatic and effortless. This is because CLT assumes that learners can draw prior knowledge from long-term memory effortlessly and compare it with the new knowledge in working memory (Sweller et al., 2019). If this assumption does not hold, as suggested by social constructivist theory in Section 2.2.2, so that activating prior knowledge is costly, then the awareness of the knowledge gap becomes another factor ignored by the CLT framework.

Another possibility less considered by the CLT model is that the learner with incorrect prior knowledge has to make choice between trusting new knowledge or old knowledge. It depends on the learner's confidence with respect to prior knowledge. Social constructivists usually name this issue 'conceptual change' (Duchi et al., 2020; Loyens et al., 2015; Nachtigall et al., 2020; Weaver et al., 2018). They argue that the failed attempts of solving the problem by the learners with prior knowledge facilitate them to change concepts. One opposing view is that learners are more likely to change their concepts through viewing than doing the experiment themselves (Renken and Nunez, 2010). Such an argument, however, does not disprove the idea that

confidence over prior knowledge is a factor influencing learning outcomes but is ignored by the traditional CLT model.

2.2.3.2 The learning phases beyond knowledge transfer

Besides influential factors ignored by the CLT model, another challenge to traditional CLT is whether there exist learning phases beyond knowledge transfer. As introduced in Section 2.2.1 and indicated by Figure 2.3, CLT assumes that the learner's prior knowledge is stored in the schema. The activities named schema automation and schema acquisition refer to activating prior knowledge from long-term memory and building transferred knowledge into long-term memory respectively. Both these schemas occur within the human information processing system, unlike knowledge transfer which occurs between the human information processing system and the outside environment. Traditional CLT assumes that knowledge transfer, schema automation and schema acquisition compete for the same resource of working memory (Sweller, 2020). The resource of working memory is allocated into intrinsic, extraneous, and germane loads. Knowledge transfer consumes intrinsic load, while schema automation and schema acquisition consume germane load. Increased extraneous load thus will reduce intrinsic and germane loads.

The underlying assumption of CLT is that knowledge transfer, schema acquisition, and schema automation belong to the same phase,[6] so they are competing for working memory resources. Such an assumption is supported by some evidence (Galy et al., 2012). However, contrary evidence also exists. Taber (2013) argued that a broad learning process contains two phases, the change of memory and the alteration of conception. Debue and Van De Leemput (2014) documented experimental evidence that there is no linear relation between intrinsic and germane loads, against the assumption that schema acquisition and schema automation belong to the same phase. A more recent experiment by O. Chen and Kalyuga (2021) suggests that the phase with limited mental sources lasts as little as one day. These authors conducted an experiment on Grade 2 students in a Singapore primary school and found that the students' cognitive load

[6] The timespan of a phase is not clearly defined. According to Van Merrienboer and Sweller (2005) and Kalyuga and Singh (2016), the duration of working memory is only around 20 seconds. However, CLT assumes the overload in the previous session will accumulate in the subsequent session. So, the phase should be the period during which cognitive loads influence each other, which should be much longer than 20 seconds.

reduced to the normal level between the immediate and delayed tests. The delayed test was on the 2nd day, indicating that a full working memory could be restored within one day.

It would be a serious challenge to CLT if schema acquisition and schema automation can belong to a different phase and do not compete for the working memory resource with knowledge transfer. PBL advocates have argued that solving problems and group discussion before instruction can help activate prior knowledge (Newman and DeCaro, 2018). The potential disadvantage of those pre-instruction activities, as argued by CLT, is that they occupy the resources of knowledge transfer. If those activities belong to different phases and consume irrelevant resources of working memory, PBL procedures only bring benefits to learning. A similar logic applies to schema acquisition. If the transferred knowledge can be built within schemas in a later phase, the activities of PBL are relevant to the learning purpose. Rittle-Johnson (2006) documented the evidence that self-explanation helps transfer. The possible missing phases in the traditional CLT model, together with the possible missing influential factors, call for modified CLT models.

2.2.4 The emerging modified CLT models

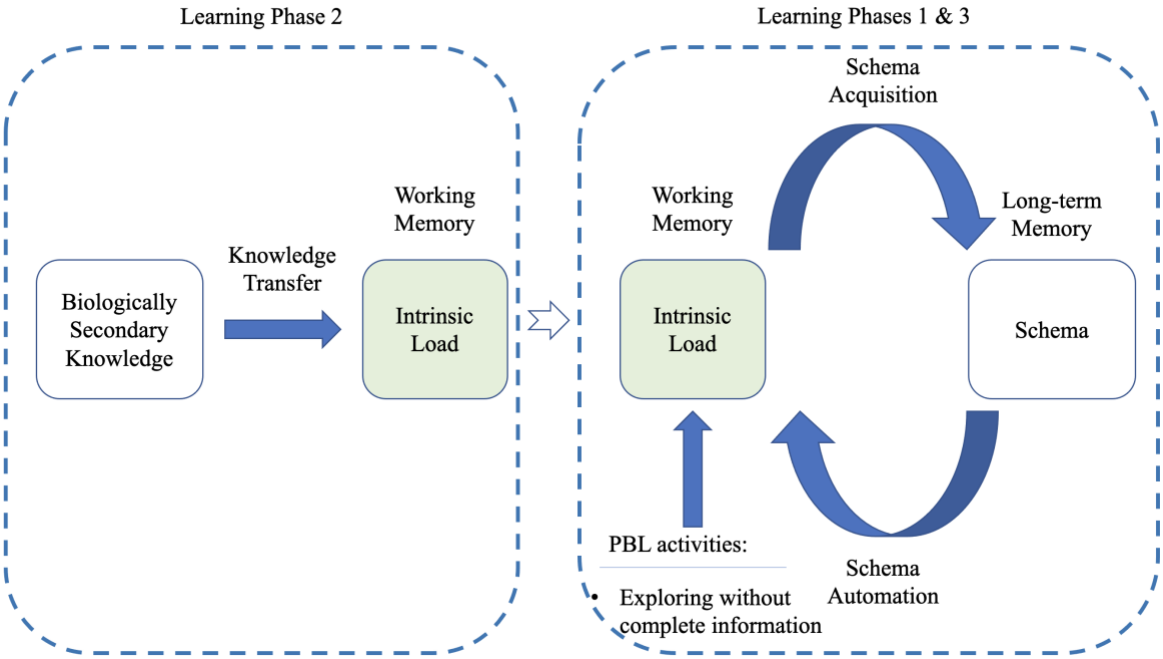
Various modified CLT models have emerged in response to the challenges discussed in Section 2.2.3, including a new integrated working memory model by Sepp et al. (2019) and an Expectancy-Value-Cost CLT model by Feldon et al. (2019). Two of those modified CLT models, which are relevant to developing the research questions of the present thesis, will now be introduced.

2.2.4.1 The interval theory view of CLT (ICLT)

The first modified CLT model is the interval theory view of CLT (ICLT hereafter) developed by Kalyuga and Singh (2016). Kalyuga and Singh (2016) argued that in complex learning, there is more than one phase. The phase described by traditional CLT, which transfers new knowledge from outside through working memory to alter long-term memory, is merely the second phase of the whole learning process. In Phase 1, the learner should activate the prior knowledge by the schema automation activity. In Phase 3, after the completion of knowledge transfer, the learner should construct the schema by schema acquisition. Whether the consumed mental resource belongs to extraneous load or intrinsic load depends on the learning goals. The activity defined by traditional CLT as the source of extraneous load could become intrinsic load, especially when

the learning task becomes complex. Kalyuga and Singh (2016) pointed out that the limitation of traditional CLT is because human cognition "has seemingly evolved to be more complex than most other natural systems" (p. 850). Figure 2.4 illustrates the conceptual framework of ICLT.

Figure 2.4: The conceptual framework of ICLT



Source: Compiled by author

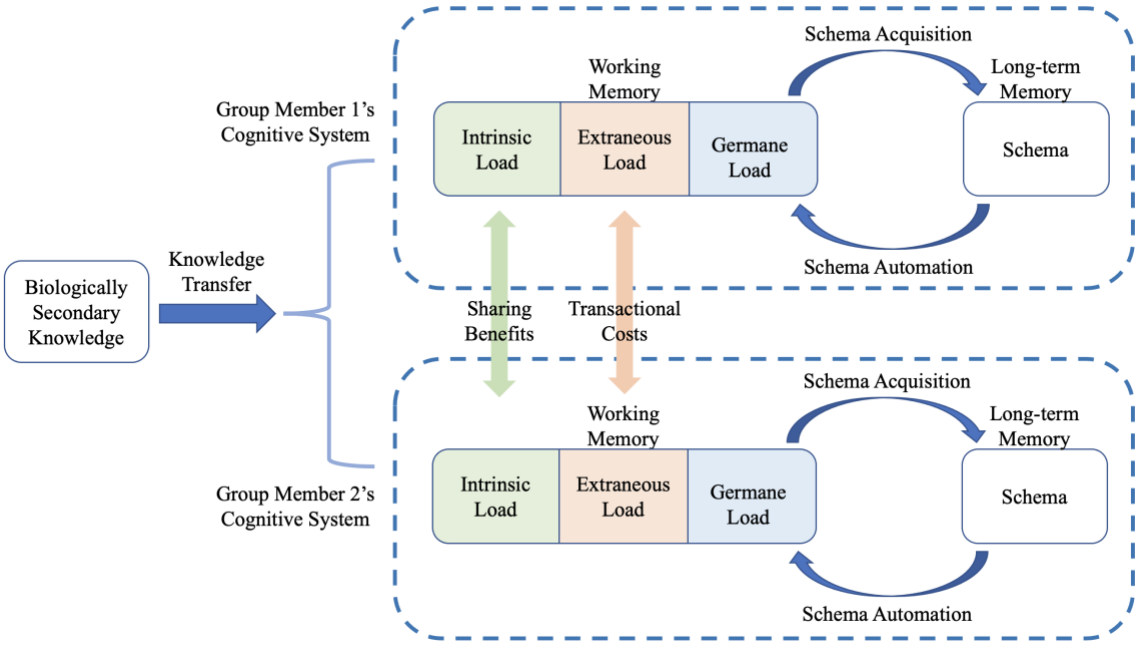
A plethora of empirical studies examining how learning task complexity impacts PBL efficacy (Blayney et al., 2016; O. Chen et al., 2020, 2021; O. Chen and Kalyuga, 2020, 2021; Likourezos and Kalyuga, 2017; Lu et al., 2020) enrich the study line of ICLT. The ICLT theory indicates that the PBL efficacy is determined by factors beyond the traditional CLT model and helps in the formulation of my research questions.

2.2.4.2 Collaborative Cognitive Load Theory (CCLT)

Another emerging modified CLT model is the collaborative CLT (CCLT hereafter) model. The traditional CLT model mainly focuses on individual learning, whereas the group discussion method, which is also one determining element of the PBL approach, has been widely regarded as an efficient way to improve learning outcomes (Zambrano et al., 2019a, 2019b). The traditional CLT model needs modification to be brought in line with that empirical evidence.

Kirschner et al. (2018) extended traditional CLT to CCLT. The way of modeling the learning process by CCLT is similar to traditional CLT but the major difference is that the former views the members in the study group as a network of cognitive systems. Figure 2.5 demonstrates the conceptual framework of Collaborative CLT.

Figure 2.5: The conceptual framework of Collaborative CLT



Source: Compiled by author

As Figure 2.5 shows, biologically secondary knowledge is transferred to the network consisting of all the group members' cognitive systems. The benefit of the grouped cognitive system is that individual members can share their working memory. Therefore, the total intrinsic load of the whole group is larger than the individual member. The group discussion in PBL thus facilitates the learning of members, especially when the learning task is complex and requires more mental resources.

There are also costs from the group discussion, which are the communication or transactional costs among group members. According to Kirschner et al. (2018), those communication activities are irrelevant to the learning goal. So, the mental resource they consume is classified as extraneous load. Zambrano et al. (2019a) argued that if the group members' prior knowledge is distributed unequally, the transactional costs will be higher.

The benefits and costs of group discussions could be affected by cultural factors. For example, Nyumba et al. (2018) stated that in a culture discouraging in-person conversation, the effects of group discussion are lower. This is another potential reason that cultural factors play a role in the efficacy of PBL. The present thesis will investigate this issue by introducing student-level cultural measures.

2.2.5 Summary of Section 2.2

Section 2.2 has briefly reviewed the cognitive foundations of traditional CLT and social constructivist learning theory. The two cognitive theories support the propositions of PBL opponents and advocates respectively. A significant drawback of social constructivist learning theory is that its model is neither standardized nor explicit. Compared to CLT, which specifies the mechanism of cognitive load, it is difficult to empirically test social constructivist learning theory. However, traditional CLT's cognitive model could be oversimplified. The challenges to traditional cognitive load theory include missing influential factors and the existence of learning phases beyond knowledge transfer.

Because of the challenges to traditional CLT, various modified CLT models have emerged. Two of those modified CLT models, the interval theory view of CLT (ICLT) and the collaborative CLT (CCLT), were introduced in Section 2.2. ICLT argues that in complex learning, the whole learning process consists of multiple phases instead of the single one of knowledge transfer identified by traditional CLT. ICLT further suggests that the standard way of distinguishing extraneous and intrinsic load is not constant across learning phases due to different learning goals.

CCLT views the members in the study group as a network of cognitive systems. The grouped cognitive regime brings both positive and negative effects to the learning outcome. These two modified CLT models, as responses to the challenges facing the traditional CLT model, can consolidate the conflict between traditional CLT and social constructivism, by integrating the cost and benefit sides of PBL into a single framework. The discussions in this section pave the way for specific propositions on PBL efficacy in the extant literature, which will be explored in Section 2.3.

2.3 Theoretical propositions and empirical evidence about PBL efficacy

The topic of PBL efficacy can be interpreted from two aspects. The first aspect is the general impact of PBL on learning outcomes. The second aspect is what factors determine the direction and the magnitude of PBL's impact. If PBL's impact is not dominated by either the positive or negative side, the second aspect is more enlightened for academic discussions and relevant to educational practitioners. This view has been called for by previous scholars (Hmelo-Silver et al., 2007; Hung, Dolmans, et al., 2019).

Based on the cognitive foundation explored in Section 2.2, traditional CLT, social constructivism, and modified CLTs all formulate theoretical propositions about PBL efficacy from the two aspects. Traditional CLT hypothesizes that PBL in general adversely influences the learning outcome, and the adverse influence is mainly caused by insufficient and late guidance in PBL. Traditional CLT believes the negative impact of PBL is more pronounced when learners have less prior knowledge or when the learning task is more complex (O. Chen et al., 2017; Sweller et al., 2011).

Conversely, social constructivists hypothesize that PBL generally is beneficial to learning. Social constructivists believe each PBL element plays a role in improving learning efficiency. However, social-constructivism-based PBL research also discusses the contextual factors which amplify the positive effects of PBL. In contrast to traditional CLT, social constructivists cover a broad range of contextual factors, including learning discipline, learner's age, forms of problem, the length of curricula, usage of digital technology, and cultural factors (Hmelo-Silver et al., 2019; Moallem, 2019; Wijnia et al., 2019).

The two modified CLTs mentioned in Section 2.2.4 can to some extent reconcile traditional CLT and social constructivism or incorporate PBL favorable propositions into the CLT framework. However, current empirical studies of ICLT still limit the contextual factors within prior knowledge and task complexity (O. Chen et al., 2020, 2021). CCLT-related studies additionally consider two new contextual factors but they mainly focus on a single element, collaborative learning (Zambrano et al., 2019b, 2019a).

To my best knowledge, so far no cognitive theories have developed formal arguments about PBL's potential synergistic effect. By defining the PBL procedure, PBL advocates implicitly

suggest that PBL is less effective if not all the key elements are present. However, those are not structured theoretical statements whereby the efficacy losses due to an incomplete PBL can be quantitatively predicted. If there exists PBL's synergistic effect, either positive or negative, PBL's overall efficacy should not equal the sum of its elements' efficacies. This section also attempts to derive the implication of PBL's potential synergistic effect from previous empirical evidence.

The present section reviews relevant empirical evidence after each part of the theoretical propositions. Included empirical papers must be either rigorous experimental studies or large-sample-based association studies. Rigorous experimental studies have a randomized and controlled pseudo-experimental design. Before-after tests or other studies not meeting rigorous experimental standards are not included by this section because they are empirically underqualified. Section 2.4 provides detailed discussion about the empirical problems. Most of the included empirical studies were published in the last 10 years, except for several influential studies published 10 to 20 years ago.

This section discusses the theoretical propositions concerning PBL efficacy and the corresponding empirical evidence in the extant literature. Following the dimension segmenting in Section 2.1, the efficacy of the three dimensions will be reviewed separately and followed by a discussion about the potential synergistic effect of PBL as a whole. The remainder of Section 2.3 is organized as follows: Section 2.3.1 presents the existing studies for the efficacy of guidance quantity; Section 2.3.2 reviews the theoretical proposition and empirical evidence about the efficacy of guidance timing or sequence; the efficacy of small-group collaborative learning is scrutinized in Section 2.3.3; Section 2.3.4 examines the synergistic efficacy of the unitary PBL approach.

2.3.1 The efficacy of guidance quantity

CLT and social constructivism distinctly develop theoretical arguments in every element of PBL. But it is guidance quantity that is the main field where the two sides have opposing predictions about PBL efficacy. The efficacy of guidance quantity thus engages considerable academic attention. I start with the first tier of propositions that how the two sides locate the optimal level of guidance quantity.

2.3.1.1 The optimal level of guidance quantity

Traditional CLT posits that the optimal level of guidance quantity is the maximum level. Mathematically speaking, traditional CLT induces a positive monotonic relationship between guidance quantity and learning efficacy. This relationship is caused by the cognitive principles depicted in Section 2.2 that copying knowledge from the instructor is always more efficient than inventing knowledge by the learner themself. The latter activity will deplete the working memory by increasing extraneous load and thus hinder the learning process. Therefore, the guidance should be at the maximum level and provide complete and accurate instruction (Sweller et al., 2019).

In contrast, social constructivism hypothesizes that the optimal level of guidance quantity can be any point except for the maximum level. Social constructivists suggest an inverted U-shaped relationship between the guidance quantity and learning efficiency. Learning efficiency will be enhanced first with increasing guidance quantity but then decline once the guidance quantity surpasses the optimal point. The traditional didactic teaching method, which provides the maximum level of guidance quantity, is usually not the optimal point (Alfieri et al., 2011; Schmidt et al., 2019).

As stated, rigorous experimental studies, and large-sample-based association studies are included in my review. However, the rigorous experimental studies usually only have a small sample of students. The large sample sources are typically based on international large-scale assessments such as OECD's (Organisation for Economic Co-operation and Development) Programme for International Student Assessment (PISA hereafter) and IEA's (International Association for the Evaluation of Educational Achievement) Trends in International Mathematics and Science Study (TIMSS hereafter) databases. The technical comparison between large-sample-based association analysis and small-sample-based experiments will be discussed in Section 2.4. Table 2.2 summarizes the results of association studies grouped by country.

Table 2.2: Association studies about the effects of reducing guidance quantity. ＋ indicates a statistically significant positive association; − indicates a statistically significant negative association; ? indicates a mixed or non-significant finding

Country	Study	Sample Size	Data Source	Knowledge acquisition	Enjoyment of science	Science self-efficacy
13 countries	Forbes et al. (2020)	74,877	PISA 2015	−		
54 countries	Cairns and Areepattamannil (2019)	170,474	PISA 2006	−		+
Australia	Kaya and Rice (2010)	3,573	TIMSS 2003	−		
	McConney et al. (2014)	4,209	PISA 2006	−	+	+
	Jiang and McComas (2015)	7,832	PISA 2006	+	−	
	Oliver et al. (2021)	14,530	PISA 2015	−		
Brazil	Hwang et al. (2018)	12,176	PISA 2012	−		
	Hwang et al. (2018)	14,360	PISA 2015	−		
Canada	Aditomo and Klieme (2020)	20,058	PISA 2015	−	−	−
	McConney et al. (2014)	5,087	PISA 2006	−	+	+
	Jiang and McComas (2015)	11,720	PISA 2006	?	−	
	Oliver et al. (2021)	20,058	PISA 2015	−		
China	Aditomo and Klieme (2020)	9,841	PISA 2015	−	?	−
	Gao et al. (2018)	457	TIMSS 2007[a]	?		
Denmark	Jiang and McComas (2015)	2,808	PISA 2006	−	−	
England	Jerrim et al. (2020)	4,361	PISA 2015[b]	+		
Finland	Lavonen and Laaksonen (2009)	4,714	PISA 2006	−		
	Jiang and McComas (2015)	2,912	PISA 2006	+	−	
	Kang and Keinonen (2018)	4,714	PISA 2006	−	?	
	Hwang et al. (2018)	5,660	PISA 2012	−		
	Hwang et al. (2018)	5,581	PISA 2015	−		
France	Hwang et al. (2018)	2,991	PISA 2012	−		

Country	Study	Sample Size	Data Source	Knowledge acquisition	Enjoyment of science	Science self-efficacy
	Hwang et al. (2018)	5,325	PISA 2015	−		
Germany	Jiang and McComas (2015)	2,317	PISA 2006	+	−	
Hong Kong	Jiang and McComas (2015)	2,355	PISA 2006	?	−	
Ireland	Jiang and McComas (2015)	2,565	PISA 2006	−	−	
	Oliver et al. (2021)	5,741	PISA 2015	−		
Italy	Jiang and McComas (2015)	13,003	PISA 2006	+	−	
Japan	Aditomo and Klieme (2020)	6,647	PISA 2015	+	?	+
	Kaya and Rice (2010)	4,250	TIMSS 2003	+		
	Jiang and McComas (2015)	4,196	PISA 2006	?	−	
Korea	Jiang and McComas (2015)	3,315	PISA 2006	?	−	
	Hwang et al. (2018)	3,348	PISA 2012	−		
	Hwang et al. (2018)	5,122	PISA 2015	−		
New Zealand	McConney et al. (2014)	1,141	PISA 2006	−	+	+
	Jiang and McComas (2015)	2,844	PISA 2006	?	−	
	Oliver et al. (2021)	4,520	PISA 2015	−		
Norway	Teig et al. (2018)	4,382	TIMSS2015	?		
	Hwang et al. (2018)	3,008	PISA 2012	−		
	Hwang et al. (2018)	5,096	PISA 2015	−		
Peru	Hwang et al. (2018)	3,817	PISA 2012	−		
	Hwang et al. (2018)	6,021	PISA 2015	−		
Qatar	Areepattamannil (2012)	5,120	PISA 2006	?	+	
	Hwang et al. (2018)	6,606	PISA 2012	−		
	Hwang et al. (2018)	9,803	PISA 2015	−		

Country	Study	Sample Size	Data Source	Knowledge acquisition	Enjoyment of science	Science self-efficacy
Scotland	Kaya and Rice (2010)	2,665	TIMSS 2003	−		
Singapore	Aditomo and Klieme (2020)	6,115	PISA 2015	−	−	−
	Kaya and Rice (2010)	6,122	TIMSS 2003	+		
	Hwang et al. (2018)	3,656	PISA 2012	−		
	Hwang et al. (2018)	5,687	PISA 2015	−		
Spain	Jiang and McComas (2015)	11,467	PISA 2006	?	−	
Switzerland	Jiang and McComas (2015)	6,283	PISA 2006	?	−	
Taiwan	Liou (2021)	7,708	PISA 2015	−	+	+
	Jiang and McComas (2015)	5,438	PISA 2006	?	?	
	Liou and Jessie Ho (2018)	4,046	TIMSS 2007	?		
	Hwang et al. (2018)	4,003	PISA 2012	−		
	Hwang et al. (2018)	7,056	PISA 2015	−		
UK	Jiang and McComas (2015)	7,911	PISA 2006	?	−	
	Oliver et al. (2021)	14,157	PISA 2015	−		
USA	Kaya and Rice (2010)	7,623	TIMSS 2003	−		
	Zhang and Li (2019)	6,503	TIMSS 2007	−		
	Jiang and McComas (2015)	3,410	PISA 2006	?	−	
	Hwang et al. (2018)	3,256	PISA 2012	−		
	Hwang et al. (2018)	5,094	PISA 2015	−		
	Oliver et al. (2021)	5,712	PISA 2015	−		

[a]This study collects the information from five middle schools in the Inner-Mongolia of China and creates a mini dataset following the standards of TIMSS 2007. [b]This study uses a dataset merged from PSIA 2015 and National Pupil Database.

Source: Compiled by author

Table 2.2 presents empirical findings about the effects using association analyses. Thanks to the two groups of international assessment databases, the association analyses covered a broad country basis. The PISA series database covers students aged 15 years (grade 11). The students from the TIMSS series database are in grade 4 or 8. The PISA database determines students' performance in scientific domains whereas the TIMSS database provides students' scores in mathematics and science. The variables measuring guidance reduction are constructs based on questioner items of PISA or TIMSS. The choice by the scholars who conducted those association analyses of PISA items to proxy for the guidance level may affect the results. See Section 2.4.2 and Appendix 3 for discussions. Besides knowledge acquisition, some studies also examine the effects of reducing guidance quantity on students' enjoyment of science and self-efficacy in science.

Despite variance across countries,[7] most results (Cairns and Areepattamannil, 2019; Liou, 2021; McConney et al., 2014) in Table 2.2 suggests that reducing guidance quantity provokes poorer scientific performance but is conducive to students' enjoyment in science and self-efficacy in science. Specifically, Cairns and Areepattamannil (2019) conducted the association analysis by pooling the PISA data of 54 countries. One study which is not included in Table 2.2 is the comprehensive technical report about PISA 2015 data by OECD (Mostafa et al., 2018). Mostafa et al. (2018) showed that for most countries the increases in inquiry-based learning, which is a proxy of less guidance quantity, are positively associated with enjoyment (Figure 3.4., p. 23) and self-efficacy (Figure 3.5., p. 24) but negatively related to scientific performance (Figure 3.8, p. 29).

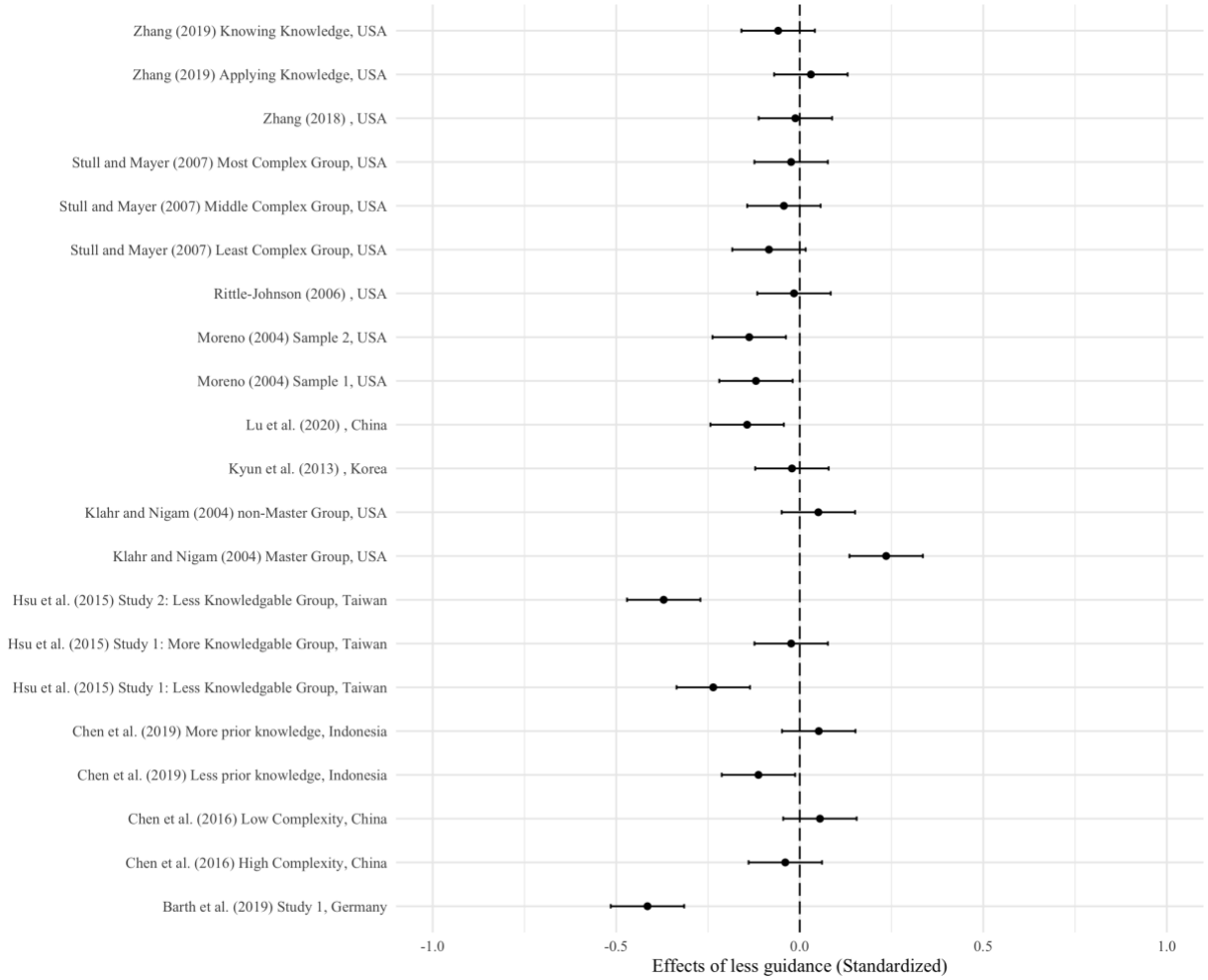
Several studies in Table 2.2 support the proposition of social constructivism that the optimal level of guidance quantity is located in the middle of the whole spectrum. Aditomo and Klieme (2020) used confirmatory factor analysis (CFA hereafter) to construct three levels of guidance. In 18 of the 20 countries investigated by Aditomo and Klieme (2020), students benefit the most from a moderate level of guidance (the two exceptions are Jordan and Kosovo). This suggests that a very large amount of guidance in learning has a detrimental effect. Teig et al. (2018)

[7] For example, the results from Singapore (Kaya and Rice, 2010), England (Jerrim et al., 2020), and Japan (Aditomo and Klieme, 2020; Kaya and Rice, 2010).

provide supportive evidence of an inverted U-shaped relationship between guidance quantity and learning outcome. By including the quadratic term of guidance quantity, Teig et al. (2018) documented that learning performance will first increase and then decrease along with reducing guidance. Given that traditional CLT holds a more aggressive proposition that guidance quantity should be at a maximum, the results of the association analyses can be considered to be more favorable for PBL advocates.

Compared to large-sample-based association analyses, there have been also many scholars investigating how guidance quantity affects learning using small sample but randomized controlled experiments. Figure 2.6 summarizes the studies that belong to this genre.

Figure 2.6: Experimental findings about the effects of less guidance



Source: Compiled by author

To facilitate comparison, Figure 2.6 presents the standardized[8] experimental results about the effects of less guidance. What is meant by guidance varies across studies. Several studies (O. Chen et al., 2016, 2019; Klahr and Nigam, 2004; Kyun et al., 2013) use worked examples as a proxy for the implementation of maximum guidance. But, as I discussed in Section 2.1.3.1, worked examples can be a scaffolding tool of PBL. Other instances of less guidance include no-instruction PBL (Barth et al., 2019), principle-absent instruction (Hsu et al., 2015), and the integrated–integrated way of teaching Chinese characters (Lu et al., 2020). For learning outcome measures, most studies[9] in Figure 2.6 use an immediate or short-term testing score. The drawbacks of only using short-term testing scores will be discussed in Section 2.4.3.

Unlike association analyses covering a number of countries, thanks to the data availability of PISA and TIMSS, most experimental studies have been conducted in the USA(Klahr and Nigam, 2004; Moreno, 2004; Rittle-Johnson, 2006; Stull and Mayer, 2007; L. Zhang, 2018, 2019). Experimental studies of the guidance quantity effect have been undertaken in Germany (Barth et al., 2019), Indonesia (O. Chen et al., 2019), Korea (Kyun et al., 2013), Taiwan (Hsu et al., 2015), and China (O. Chen et al., 2016; S. Gao et al., 2018; Lu et al., 2020). Studies outside the USA tend to achieve more significant results. One possible explanation is that the experiments out of the USA are less likely to measure learning outcomes with transfer or delayed test design. For example, S. Gao et al. (2018), Barth et al. (2019), and study 1 of Hsu et al. (2015) all use the score of retention tests immediately after learning. O. Chen and Kalyuga (2021) suggested that the inquiry-based learning approach has delayed and transferred advantages in learning efficacy.

For the same reason that I gave in Section 2.1.3.1, experimental studies showing the adverse effects of less guidance cannot indicate the superiority of the traditional CLT. Social constructivism also predicts a positive correlation between guidance quantity and learning

[8] The magnitudes of effects are standardized so all the studies have the same standard error. It inevitably suffers limitations as it will understate the absolute value of the mean effect if the association analysis inherently has a greater variance of guidance efficacy. It does not change the nature of testing since a larger variance of guidance efficacy means it is more difficult to reject the null hypothesis. However, it is worth noting that the standardized value of incremental effect here should not be inferred as the magnitude level of the incremental effect of treatment. The technical details of standardization can be found in the Appendix 2.

[9] Exceptions are Klahr and Nigam (2004), Rittle-Johnson (2006), and Hsu et al. (2015) who carried out tests delayed by 5 to 7 days.

outcomes when the guidance quantity is below the optimal level. As Kapur (2016) said, rejecting PBL without any instruction "does not logically mean the maximal provision of the guidance is the most effective solution" (p. 291). Rather, the observed negative impacts of the increased guidance quantity in some experiments can validate the proposition of PBL advocates. For instance, some CLT-based studies like O. Chen et al. (2016) found a positive impact of reducing guidance in a subgroup of the experiments. The reversal effects cannot be explained by a traditional CLT model[10] because in a traditional CLT model, if the working memory is not overloaded, the negative impacts of less guidance on learning efficiency will only disappear but not be converted into positive impacts. O. Chen et al. (2016) attributed their findings to the generation effect to recoup the loss of learning efficiency, which means that the traditional CLT is incomplete and should be modified.

2.3.1.2 The contextual factors affecting guidance efficacy

Given the uncertain line between the maximum-level explicit instruction and optimal scaffolding in the definition of PBL discussed in Section 2.1.3.1, it is more meaningful to investigate the contextual factors determining the optimal level of guidance (Hung, Dolmans, et al., 2019). Traditional CLT postulates that the negative effects of not choosing maximum guidance quantity will be less pronounced if the learners have more prior knowledge or the learning task complexity is lower. CLT believers have named the above two contextual propositions as Expertise reversal effect and Element interactivity effect, which happens when the intrinsic load is small so that increased extraneous load cannot deplete working memory (O. Chen et al., 2017).

PBL advocates also agree that those two factors can affect the guidance efficacy. Yet their argument is from the benefit side of withholding information. For instance, Richey and Nokes-Malach (2013) claimed that withholding instructional explanations fosters students' constructive cognitive activities. As shown in Figure 2.6, most recent experimental studies investigate the contextual effects of those two factors. In particular, Klahr and Nigam (2004), Hsu et al. (2015),

[10] CLT scholars actually assert that the cognitive load effect will be reversed for the expert learner. For example, Sweller et al. (2011) said, "That advantage disappeared or even reversed for higher prior knowledge learners" (p. 210). This argument, also named as expertise reversal effect, can not be explained by John Swell's CLT model and mainly reflects Slava Kalyuga's modified CLT thoughts.

and O. Chen et al. (2019) found that less experienced learners need more guidance or are less suitable for the PBL approach.

The evidence for learning task complexity is relatively mixed. Stull and Mayer (2007) found that the negative effects of reducing guidance are pronounced when the learning task is less complex. This is contrary to the prediction of traditional CLT which hypothesizes that for the more complex task, guidance is more beneficial to lessen learners' extraneous load. O. Chen et al. (2016) showed evidence suggesting the existence of the Element interactivity effect predicted by traditional CLT. But, as discussed in Section 2.3.1.1, the positive impacts of guidance quantity on learning outcome when the element interactivity is low demonstrated by O. Chen et al. (2016) indicate the traditional CLT cannot be the full picture. The benefit side of less guidance cannot be accounted for by the traditional CLT model.[11]

The benefit side of less guidance should be further explored by including more contextual factors. Social constructivism and modified CLTs both argue that there are more factors affecting the efficacy of less guidance, either through a completely different cognitive model or an extended cognitive model. The potential influential factors include the instructor's background (Leary et al., 2013), curriculum-wide implementation of PBL (Dolmans et al., 2016), and the application of digital scaffolding techniques (Kim et al., 2018). Results from association analyses suggest that students' grades and learning tasks may also affect guidance efficacy. For example, Kaya and Rice (2010) and Aditomo and Klieme (2020) found opposing results for Singapore. Those two studies differ in students' grades and learning tasks. Kaya and Rice (2010)'s study was based on TIMSS 2003 where students are in grade 4 and testing is for the mathematics and scientific domains. Aditomo and Klieme (2020)'s study was based on PISA 2015 where students are in grade 11 and testing is for scientific domains only.

Cultural variance could be another potential influential factor. Cross-country studies such as Cairns and Areepattamannil (2019), Forbes et al. (2020), and Aditomo and Klieme (2020), all illustrate the substantial variation in the pattern of guidance efficacy across countries. Cairns and Areepattamannil (2019) showed that 43% of the variance in learning outcome is explained by

[11] The similar situation occurs for the tests on expertise effect when the more experienced subgroups in Klahr and Nigam (2004) and O. Chen et al. (2019) exhibit better learning performance with less guidance.

unobserved country-level factors, compared with the 6% explained by student-level factors. Forbes et al. (2020) documented that Korean students are less likely to argue about science questions or draw conclusions, concluding that more factors such as cultural norms affect pedagogical practices.

2.3.2 The efficacy of late guidance

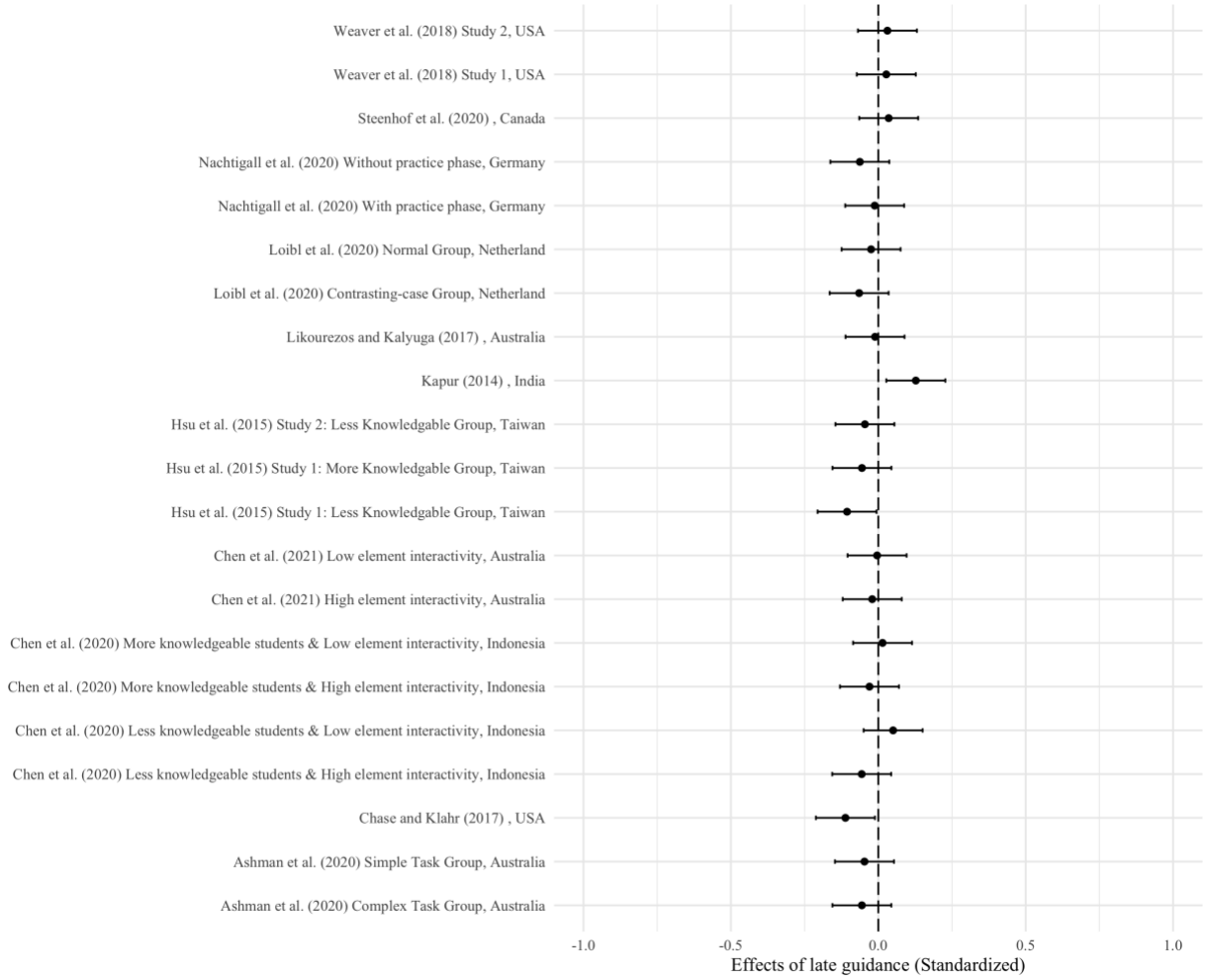
Similar to guidance quantity, the efficacy of late guidance consists of two tiers of propositions: general effects, and contextual effects. For the first-tier proposition, traditional CLT predicts that presenting the problem before instruction, or late guidance, also generates extraneous load and impairs learning efficacy (Sweller, 2020). Conversely, social constructivists believe that an initial failed attempt to solve the problem can be conducive to learners' consequent learning. Social constructivists refer to the early, but unsuccessful, problem-solving attempts as productive failures in learning (Kapur, 2012, 2014, 2016). The modified CLTs, particularly ICLT, combine both the theory of productive failure and the theory of inefficient failure increasing cognitive load (Kalyuga and Singh, 2016).

For the second-tier proposition, traditional CLT argues that the learner's prior knowledge and learning task complexity affects the efficacy of guidance timing. Social constructivism suggests that there are more contextual factors. For example, Loibl et al. (2017) reviewed previous experiments and concluded that the type of scaffolding, such as contrasting cases or building instruction, can improve the efficacy of late guidance. A more recent meta-analysis by Sinha and Kapur (2021) found that the efficacy of late guidance is stronger with high fidelity to the principles of Productive Failure. Holmes et al. (2014) argued that scaffolding improves the productivity of failure. Tawfik et al. (2015) suggested that "using a failure-based strategy may be more beneficial for diagnosis-solution problems when compared with design problems" (p. 991). In ICLT, the failure of solving the problem with students' prior knowledge forms a necessary step to prepare mental resources before the learning phase. Therefore, ICLT's theoretical framework should allow for more contextual factors. This is borne out by the review of O. Chen and Kalyuga (2020), which concluded that in addition to element interactivity and prior knowledge, the type of knowledge also determines the best guidance sequence.

The studies about the efficacy of late guidance are all experiment-based since neither PISA nor TIMSS series databases include a variable about guidance timing. Unlike guidance quantity

which suffers the problem of inconsistent measure, the timing of guidance is a dichotomy option issue: problem-instruction, or instruction-problem sequence. Similar to illustrating the effects of less guidance (Figure 2.6), I plot a diagram in Figure 2.7 to summarize the effects of late guidance.

Figure 2.7: Experimental findings about the effects of late guidance



Source: Compiled by author

The construction process of Figure 2.7 is the same as in Figure 2.6, where the effect size of late guidance is standardized.[12] Compared to guidance quantity efficacy experiments concentrating in

[12] The limited disclosure of covariance between two samples in many studies may lead to overestimated standard errors, as shown in Figure 2.7. Consequently, even if the line x=0 falls within the standard error bar, the original study could still report significant results. See Appendix 2 for technical details.

USA, the recent experiments about guidance timing are more likely to have been conducted in the rest of the world. The experimental evidence comes from Australia (Ashman et al., 2020; O. Chen et al., 2021; Likourezos and Kalyuga, 2017), Indonesia (O. Chen et al., 2020), Netherlands (Loibl et al., 2020), India (Kapur, 2014), Germany (Nachtigall et al., 2020), Canada (Steenhof et al., 2020), and Taiwan (Hsu et al., 2015). Chase and Klahr (2017), Weaver et al. (2018), and Matlen and Klahr (2013) report on studies that were carried out in the USA. Five experiments (Ashman et al., 2020; Chase and Klahr, 2017; Hsu et al., 2015; Matlen and Klahr, 2013; Steenhof et al., 2020) examined delayed learning effects. In the other studies, assessment of students' learning was undertaken within one hour of the tuition, so the learning outcome is measured by an immediate testing score.

The conclusion about the first-tier effects of guidance timing is still inconclusive, according to findings in Figure 2.7. Hsu et al. (2015), Chase and Klahr (2017), Loibl et al. (2020), and Ashman et al. (2020) found evidence in line with the proposition of traditional CLT that late guidance may have detrimental effects on learning. The results of all those studies, except for Loibl et al. (2020), are robust to tests delayed by 6 to 14 days. One subgroup in O. Chen et al. (2020) demonstrates the positive effect of late guidance, suggesting the productive side of failure dominates when the element interactivity is low. The positive effect of late guidance is also found in the studies of Kapur (2014), Weaver et al. (2018), and Steenhof et al. (2020).

Matlen and Klahr (2013), which also examined the effects of guidance timing, is not included in Figure 2.7 because it does not disclose the mean and standard errors for each subgroup. However, this USA-based experiment showed that the sequence of guidance did not affect the learning performance. Using the Australian experimental environment, Likourezos and Kalyuga (2017) also concluded that the different learning paths lead to a similar learning outcome. A meta-analysis by Sinha and Kapur (2021) found although late guidance has no significant benefits for learning procedural knowledge, it is more helpful for learning conceptual knowledge. The findings of Matlen and Klahr (2013), Likourezos and Kalyuga (2017), and Sinha and Kapur (2021), together with findings in Figure 2.7, render the general effect of late guidance still inconclusive.

For the two contextual factors of traditional CLT, the effects of learners' prior knowledge are not consistent in extant studies. Hsu et al. (2015) showed that the negative impacts of late guidance

cease for the more knowledgeable students, which were grade 11 and 12 students in their experiment, compared to the less knowledgeable group of grade 10 students. However, the experiment conducted by O. Chen et al. (2020) in Indonesia shows that the gap between late and early guidance is widened for less knowledgeable students. O. Chen et al. (2020)'s findings oppose the hypothesis of traditional CLT in terms of learners' prior knowledge. On the other hand, the increased learning task complexity enhanced the advantage of early guidance in the experiments of Ashman et al. (2020), O. Chen et al. (2020), and O. Chen et al. (2021). It is worth noting that learning task complexity is not equal to task difficulty; O. Chen et al. (2017) distinguished the two concepts. The extant research usually uses element interactivity as the proxy for learning task complexity.

Despite prior knowledge and task complexity, PBL advocates also suggest that the knowledge type matters for the efficacy of late guidance. From their viewpoint, early guidance mainly helps learners to acquire basic knowledge. For knowledge that needs conceptual change (Kapur, 2016) or transfer (Loibl et al., 2017), late guidance is believed to result in more productive learning outcomes. Kapur (2016) even claimed that for conceptual change knowledge, early guidance leads to unproductive success. Chase and Klahr (2017) similarly suggested that the effectiveness of guidance timing is contingent on what kind of knowledge is taught. In addition to knowledge type, Holmes et al. (2014) argued that the subsequent scaffolding is a critical determinant of whether the initial failure is productive or not. There is no evidence indicating that cultural factors play a role in the efficacy of guidance sequence, due to the lack of cross-country studies.

2.3.3 The efficacy of small-group collaborative learning

Traditional CLT did not bring in the proposition for collaborative learning since it considers the learning as an individual task (Sweller et al., 2011) As discussed in Section 2.2.4, CCLT extends the single cognitive system to a network of cognitive systems. While collaborative learning in groups can enhance working memory by sharing cognitive resources, it also creates additional extraneous load due to communication among group members. Consequently, determining the efficacy of collaborative learning by CCLT is not straightforward, and there is no single optimal solution. Instead, the effectiveness of small-group collaborative learning depends on the trade-off between its benefits and costs, which is affected by contextual factors. In particular, CCLT

suggests two new contextual factors: prior collaborative experience and information distribution among group members (Kirschner et al., 2018).

Social constructivists have a similar proposition that the efficacy of small-group collaborative learning depends on contextual factors. They implicitly assume that small-group collaborative learning can provide a positive effect and they enumerate more contextual factors such as autonomy, group climate, and culture (Fonteijn and Dolmans, 2019; Gillies, 2016; Slavin, 2014). Culture could be an important contextual factor influencing the efficacy of small-group collaborative learning. Nyumba et al. (2018) argued that in a culture discouraging verbal conversation, the effects of group discussion are reduced. Robinson et al. (2015) suggested that the PBL group is less effective for groups with diverse cultures. Asian culture is also regarded as not encouraging direct and open communication (Choon-Eng Gwee, 2008). Accordingly, collaborative learning is less effective in Nepal (Holen et al., 2015), India (Nanda and Manjunatha, 2013), and Japan (Imafuku et al., 2014).

The first-tier efficacy of collaborative learning is mixed. Although experimental results such as those of Zambrano et al. (2019b) indicate overall positive effects of group learning, the experiment by Weaver et al. (2018) found that collaborative learning does not contribute to students' learning compared to individual learning. The findings from association analyses suggest that collaborative learning, at least in certain circumstances, has negative impacts. For example, there is one item in the PISA 2006 database named "There is a class debate or discussion." A negative association between class debate/discussion and learning outcome has been found (Lavonen and Laaksonen, 2009). Although the group size of class debate is not known and may exceed the upper boundary suggested by PBL practitioners, this finding at least provides evidence that collaborative learning could be associated with worse learning outcomes in certain situations. In Zambrano et al. (2019b), the knowledgeable individual learner has an even better score in the delayed test, contrary to the decaying effect in other experiments. It would be interesting to explore the underlying mechanism if it is not driven by design biases. Zambrano et al. (2019b) also showed that less knowledgeable learners, who achieve worse learning performance, have lower cognitive loads. This finding is also in conflict with the propositions of traditional CLT.

The results about the second-tier efficacy of collaborative learning are congruent with CCLT. The experimental evidence first suggests that the original two contextual factors of traditional CLT still influence the efficacy of group learning. Zambrano et al. (2019b) conducted randomly-controlled experiments in Ecuador. They partitioned the students into four subgroups by collaborative learning and prior knowledge. Their findings showed that collaborative learning is more useful when the learners are less knowledgeable. Several recent experimental studies support the arguments of CCLT about the efficacy of small-group collaborative learning. Zambrano et al. (2019a) found that if the group members' prior knowledge is distributed unequally, the learning is more efficient. Zambrano et al. (2019a) also documented the positive effect of the learners' prior collaborative experience, which is different from learners' prior knowledge. In summary, the two new contextual factors suggested by CCLT are both supported by Zambrano et al. (2019a).

2.3.4 The synergistic efficacy of PBL

It can be argued that the PBL approach is synergistic, which means the PBL as a whole contributes more than the sum of its three dimensions. But, unfortunately, neither the supporters nor the opponents of PBL have developed a theoretical basis for the synergistic effect of PBL. As mentioned in Section 2.1, when PBL advocates define the standard procedure, they combine all the elements together. The group discussion should revolve around the problem before the instruction is given. Therefore, all the key elements or dimensions of PBL should be the necessary conditions for an effective PBL approach, suggesting a positive synergistic effect.

Ideally, the synergistic effect can be tested by comparing the PBL's overall effects and the sum of the effects of its elements. But it is difficult to find a reasonable way to sum the amounts of the individual effect; a discussion about the relationship between testing scores and learning output is required. Alternatively, the synergistic effect can be indirectly rejected by the interactive experimental design. For example, if we observe that all the individual elements can improve the learning outcome, but with the same experimental environment the overall PBL method does not contribute to better performance than the sum of the individual elements, we can reject the null hypothesis of positive synergistic efficacy. Matlen and Klahr (2013) did much what I described above, except that they only tested the interactive effects between less guidance and late guidance. Matlen and Klahr (2013)'s findings show the subgroup with both less

guidance and late guidance ended up with the worst learning outcome. Therefore, the results of Matlen and Klahr (2013) can be regarded as evidence opposing the existence of the positive synergistic efficacy of PBL.

There is another way to draw implications from the extant literature about the synergistic efficacy of the PBL approach. Given the inclusive first-tier effects of less guidance (see Section 2.3.1), late guidance (see Section 2.3.2), and collaborative thinking (see Section 2.3.3), if the experimental evidence is generally more favorable for PBL advocates, we can attribute that to a positive synergistic efficacy. Several meta-analyses of PBL efficacy, such as Dagyar and Demirel (2015), Wijnia et al. (2017), Merritt et al. (2017), Juandi and Tamur (2021), and X. Gao et al. (2022), concur with such a hypothesis.

However, most studies using PBL as treatment are not empirically qualified as reliable evidence. An earlier review by Minner et al. (2010) (Table 3, p. 485) found that 78% of the studies about student-centered learning method do not have a meaningful control group. I reviewed more recent studies and found the empirical problems remain. For example, Penjvini and Shahsawari (2013) and Westhues et al. (2014) both claimed they randomly assigned students into PBL and control groups, but the students' pretest scores differed significantly across the two subgroups. The empirical caveats of Penjvini and Shahsawari (2013) and Westhues et al. (2014) suggest the positive overall effect of PBL could be attributed to inherent differences instead of PBL. Even if the students are randomly assigned, as in Kazemi and Ghoraishi (2012), Ajai et al. (2013), and Imanieh et al. (2014), their teachers for the PBL and control groups are not randomly assigned. Some before-after studies like study 2 of Barth et al. (2019) even have no control group. Using the overall PBL method as the treatment variable encounters empirical challenges due to its intrinsic characteristics; Section 2.4 provides more detailed discussions about this issue.

Removing studies with empirical problems, there are few PBL studies left. Their results, furthermore, cannot reject the null hypothesis of negative or zero synergistic efficacy of PBL. De Witte and Rogge (2016) is one PBL-efficacy study with rigorous empirical design. The results of its experiment based on secondary students in Belgium, however, suggest that the traditional teaching method leads to a significantly better learning outcome than PBL. Therefore, the current findings about the overall PBL effect do not provide a more encouraging conclusion than the

sum of the individual PBL elements. Thus, it is difficult to infer positive synergistic PBL efficacy from the existing literature.

2.3.5 Summary of Section 2.3

This section reviews the theoretical propositions and empirical evidence of PBL efficacy. For the first-tier effect of PBL elements, the current empirical evidence is still mixed. In contrast to the traditional CLT interpretation (L. Zhang et al., 2022), the present thesis suggests inferring the evidence based on a commonly accepted definition. Therefore, worked examples should be considered as one kind of scaffolding. The general negative relationship between guidance quantity and learning outcome cannot be inferred as strong evidence against PBL. Rather, the observed inverted U-shaped relationship in several studies can support the proposition of PBL advocates. The general effect of guidance timing and collaborative thinking is also inconclusive.

The large-sample-based association studies demonstrate a tremendous variance in learning patterns and PBL element efficacy across countries, especially between east Asian and western countries. The current experimental studies about guidance quantity mainly focus on the USA and pay less attention to the rest of the world. The pattern of fewer empirical studies in other countries found in the present literature review is in line with the call by Hung, Dolmans, et al. (2019) that more studies "from non-western cultural contexts are needed to expand the spectrum of PBL literature" (p. 952).

For the second-tier or contextual effect of PBL elements, the existing results are more favorable to modified CLTs. Both learners' prior knowledge and learning task complexity determine the efficacy of PBL elements. However, the observed reversal effect is in line with ICLT instead of traditional CLT. The extant empirical evidence also shows that the information distribution density among group members can affect the learning performance of collaborating thinking, congruent with the proposition of CCLT. Although the framework of modified CLTs allows for more contextual factors, the current empirical studies about contextual factors of PBL effects are still confined within the territory of traditional CLT. More contextual factors should be tested.

This section also discusses the potential synergistic efficacy of PBL, which is an issue less covered by both PBL advocates and opponents. Due to prevalent empirical pitfalls in experiments investigating the overall PBL effect, it is difficult to draw meaningful implications

about the synergistic efficacy of PBL. A rigorous empirical design mitigating those pitfalls is thus needed.

2.4 The pitfalls of the empirical strategies used by PBL-related research

The divergent theoretical propositions call for empirical work to test them. However, pitfalls are prevalent in PBL's empirical studies. Minner et al. (2010) surveyed the research methodology for the topic of inquiry-based science learning from 1984 to 2002. Among the 138 studies they reviewed, only 30 were designed to the experimental standard (a term used by Minner et al. to refer to a benchmark or reference point used to assess the accuracy or effectiveness of an experiment). A more recent review by L. Zhang et al. (2022) asserted that empirically dubious studies persist in the field of examining PBL efficacy, and so mislead policymakers. L. Zhang et al. (2022) claimed that "It is obvious that selecting a different category of research will result in different implications for educational practice" (p. 1170). Correctly assessing empirical quality is not only helpful for understanding of current empirical evidence related to PBL but also necessary for formulating the research design of the present thesis in Chapter 4.

This section explores those empirical pitfalls by research methodological category. Particularly, Section 2.4.1 refers to the problem of the pre-post non-experimental comparison without a control group, Section 2.4.2 evaluates the technical challenges in large-sample-based association PBL studies, and Section 2.4.3 discusses the common issues in PBL's experimental design.

2.4.1 The pre-post non-experimental test

The first category of PBL research, which compares the learning outcome before and after the implementation of the PBL approach, can be labeled as a pre-post non-experimental test. The term non-experimental means that there is no control group of students receiving the traditional instructions. The change measured by the pre-post design includes not only PBL's efficacy but also other confounding effects, such as increases in learning time, the pivot of learning attitude, the change in teacher engagement, and even students' accumulative knowledge acquisition (L. Zhang et al., 2022). Along with time, students may improve without any external instructions, or the students' learning performance could improve even more with traditional teaching. Therefore, it is difficult to infer the empirical results from the pre-post non-experimental test without excluding those confounding factors.

Non-experimental pre-post test studies are typically published in an earlier period when the PBL pilot program is initially launched. In the period reviewed by Minner et al. (2010), around 53% of studies have no control group. Even if those tests show the learning outcome has been improved, it could be attributed to other time-varying confounding factors.

Although fewer publications now use the pre-post non-experimental test design compared to the era of Minner et al. (2010), they can still be found in recent years. For example, study 2 of Barth et al. (2019) compared the knowledge scores of Education-Master-program students in a German university before and after a change in teaching method. Frederiksen (2018) is another example of the pre-post non-experimental test which investigated the effects of the PBL approach on a discourse teaching program. Two Indonesia-based experiments of Simamora et al. (2017) and Khoiriyah and Husamah (2018), which examined the mathematical problem-solving skills of only one group of students with PBL, are also pre-post non-experimental tests. The individual results of pre-post non-experimental tests have limited empirical value in the ongoing debates about PBL efficacy. However, when these results are combined and analyzed through a meta-analysis, they may have greater empirical value. This is because the biases due to unobserved confounding factors that may be present in individual studies could potentially be offset when the findings are combined and analyzed in a meta-analysis.

2.4.2 Large-sample-based association analyses

In Section 2.3.1.1, I mentioned the large-sample-based association analyses. Those are regression analyses using variables from PISA or TIMSS databases. The independent variable is the qualitative teaching activity measure derived from questionnaires to students or teachers. The dependent variable is the scientific literacy score from the standard assessment of PISA and TIMSS. The regressions are usually at the student level. So, most of these regression analyses used the hierarchical linear modeling technique because students are clustered within the classroom, school, and country. A few studies of this category employed GLS (generalized least squares) analysis for sample survey data (Lavonen and Laaksonen, 2009). Some studies (S. Gao et al., 2018; Hwang et al., 2018; Kang and Keinonen, 2018; Oliver et al., 2021; Teig et al., 2018) use normal OLS (ordinary least squares) studies, which may be subject to heteroskedasticity issue (Wooldridge, 2010).

Compared to small-sample-based experimental tests, the findings from the large-sample-based association analyses have less bias due to specific samples. The public objective data of PISA or TIMSS are also less likely to be manipulated. The large-sample feature allows regressions to control for more confounding variables because of the mass degree of freedom in statistics compared to small-sampled experiments. However, large-sample-based association analyses suffer from several empirical pitfalls.

The first pitfall of large-sample-based association analyses is the difficulty of inferring causality. Unlike pseudo-experimental or experimental designs, association or correlational tests cannot realistically produce causality inference. The structure of PISA or TIMSS is clustered cross-sectional data. We cannot examine the change of both independent and dependent variables over time. All we derive from the regression is the static association between guidance quantity and students' learning performance.

The reversed causality could lead to contrary inference for the negative association. Students who do not learn well can improve more from the PBL approach. A competing explanation can be supported by certain results in the association analyses. For example, in Forbes et al. (2020)'s association analysis based on 13 countries, all the teaching activities including the two teacher-directed items (ST098Q06TA and ST098Q09TA in PISA 2015, see detailed explanation in Appendix 3) are negatively associated with the learning outcome. The reversed causality from learning performance to teaching practice makes more sense here, otherwise, we should conclude that any form of teaching activity deters students' learning.

The second pitfall is that both PISA and TIMSS provide limited fixed variables, narrowing the scope and depth of any analysis. As mentioned in Section 2.3.2, the efficacy of guidance sequence cannot be examined with PISA or TIMSS data. Furthermore, with large-sample-based association analyses, we cannot check the mediation effect of cognitive load during different learning phases, which is a powerful mechanism to distinguish the propositions of traditional CLT and social constructivism. The lack of a flexible experimental environment is the disadvantage of the large-sample-based association analyses.

The third pitfall is that the measure of instruction practice is subjective, as perceived by students or teachers. As shown in Appendix 3, PISA and TIMSS databases collect the scaled answers

from students and teachers from a questionnaire about instructional practice in the classroom. Students or teachers choose from 1 to 4 to indicate the frequency of each instruction-related item. Previous studies have noticed the potential discrepancy between the perceived data and the actual teaching practice. For instance, Liou (2021) argued that "the students' perceived higher frequency of teacher-directed instructional practices did not represent that the learning contexts are less autonomous" (p. 328). Students who had previously experienced more PBL approaches could be more likely to perceive a higher frequency of instruction in the present classroom. Thus, their better learning performance might reflect the lasting effects of their previous PBL.

The last but not least pitfall is that there is no standard way to construct guidance quantity from PISA or TIMSS data. PISA and TIMSS are both a series of data, updating the item definitions yearly. Liou (2021) claimed that there are fewer TIMSS-based studies than PISA-based ones. TIMSS has no official measure of inquiry-based learning; scholars choose items according to their research purpose. Liou (2021) suggested that TIMSS-based studies are more flexible in constructing teaching-practice measures. For example, using TIMSS data, Kaya and Rice (2010) selected five activities as proxies for less guidance, while L. Zhang and Li (2019) choose only three.

The 2006 version of PISA is similar to TIMSS. OECD (2009) identified 17 items related to teaching practice in the classroom. Scholars have selected certain items among the 17 to construct their measure as a proxy for guidance quantity. I surveyed the studies based on PISA 2006 and listed which items they chose in Table 2.3. The detailed explanation of the PISA Item Code in the table can be found in Appendix 3.

Table 2.3: The PISA2006 items chosen by studies

PISA Item Code	Areepattamannil (2012)	Cairns and Areepattamannil (2019)	McConney et al. (2014)	Lavonen and Laaksonen (2009)	Jiang and McComas (2015)	Kang and Keinonen (2018)
SCINTACT	✓					
ST34Q01		✓	✓			
ST34Q05						
ST34Q09						

PISA Item Code	Areepattamannil (2012)	Cairns and Areepattamannil (2019)	McConney et al. (2014)	Lavonen and Laaksonen (2009)	Jiang and McComas (2015)	Kang and Keinonen (2018)
ST34Q13						
SCHANDS	✓					
ST34Q02					✓	
ST34Q03		✓	✓	✓	✓	✓
ST34Q06		✓	✓	✓	✓	
ST34Q14					✓	
SCINVEST	✓					
ST34Q08		✓	✓	✓	✓	✓
ST34Q11		✓	✓	✓	✓	✓
ST34Q16		✓	✓	✓	✓	✓
SCAPPLY	✓					
ST34Q07						
ST34Q12						
ST34Q15						
ST34Q17						

Source: Compiled by author

Table 2.3 indicates that there is no standard way of constructing a measure of less guidance in learning. In particular, Kang and Keinonen (2018) creates the measure using CFA. The results of the association analyses are affected by the choice of those items. Including items SCINTACT (composite index of Q01, Q05, Q09, and Q13) and SCAPPLY (composite index of Q07, Q12, Q15, and Q17) contributes to a positive relation between less guidance and scientific achievement (Areepattamannil, 2012). However, for studies (Cairns and Areepattamannil, 2019; Lavonen and Laaksonen, 2009; McConney et al., 2014) constructing the independent variable only using items SCHANDS and SCINVEST or their sub-items, the associations between less guidance and scientific achievement are more likely to be negative.

In the 2015 version of PISA, an official measure of inquiry-based teaching is provided. OECD (2017) defined a special item named IBTEACH to represent a teaching method with less guidance. The item named TDTEACH represents a teaching method with more explicit guidance. However, the official measure of PISA 2015 has been challenged by some scholars, with Jerrim et al. (2020) and Liou (2021) arguing that two sub-items of IBTEACH are not relevant to less guidance instruction and therefore excluding them from their analysis. I surveyed the studies based on PISA 2015 and listed which items they chose in Table 2.4. In particular, Lau and Lam (2017) and Aditomo and Klieme (2020) create their measures using CFA. (The detailed explanation of the PISA Item Code in the table can again be found in Appendix 3.)

Table 2.4: The PISA2015 items chosen by studies

PISA Item Code	Jerrim et al. (2020)	Liou (2021)	Forbes et al. (2020)	Aditomo and Klieme (2020)	Oliver et al. (2021)	Lau and Lam (2017)	Hwang et al. (2018)
IBTEACH					✓		
ST098Q01 TA	✓	✓	✓	✓			✓
ST098Q02 TA	✓	✓	✓	✓			✓
ST098Q03 NA	✓	✓	✓	✓		✓	✓
ST098Q05 TA	✓	✓	✓	✓			✓
ST098Q06 TA			✓	✓			✓
ST098Q07 TA	✓	✓	✓	✓		✓	✓
ST098Q08 NA	✓	✓	✓	✓		✓	✓
ST098Q09 TA			✓	✓			✓
ST098Q10 NA	✓	✓	✓	✓			
TDTEACH					✓	✓	
ST103Q01 NA		✓		✓			
ST103Q03 NA				✓			

PISA Item Code	Jerrim et al. (2020)	Liou (2021)	Forbes et al. (2020)	Aditomo and Klieme (2020)	Oliver et al. (2021)	Lau and Lam (2017)	Hwang et al. (2018)
ST103Q08 NA		✓		✓			
ST103Q11 NA		✓		✓			

Source: Compiled by author

In summary, although PISA and TIMSS are standardized databases, there still exists room for scholars to choose the items to construct their variable of interest. Prior empirical findings suggest that the choice of items matters for the relationship found between less guidance and learning performance. The non-standard measure issue, together with the possible biases from students' perceptions and ambiguous causality, add to the empirical pitfalls of large-sample-based association analyses.

2.4.3 Experimental and pseudo-experimental tests

The best way conducive to causality inference is by designing an experimental or pseudo-experimental test. However, the experimental feature means an environment where only one factor of interest is changed and all other confounding factors should be statistically the same between treatment and control groups. Such an experimental environment necessitates delicate designs including randomized grouping and well-defined treatment factors. For experiments about PBL efficacy, violations of these principles are prevalent. Table 2.5 summarizes experiments with PBL as the treatment factor in recent years.[13]

Table 2.5: Empirical design issues in PBL experimental tests

Study	Duration	Randomly assigned students	Randomly assigned teacher	Identical learning material	PBL procedure is stated clearly	Effect size
De Witte and Rogge (2016)	less than 1 hour	Yes	Yes	Yes	Yes	-2.26

[13] I searched the keyword "Problem-based learning" on ScienceDirect, JSTOR, and Google scholar databases. I also refer to the reference lists of review or meta-analysis studies including Merritt et al. (2017), Juandi and Tamur (2021), and X. Gao et al. (2022). I first stripped the studies without a control group. The experiments not comparing PBL and traditional lecturing methods are also excluded. This process ends up with 12 studies. There could be still experimental studies missed by this list. However, the papers in Table 2.5 should be more than representative of the recent experimental studies with PBL as a treatment variable.

Study	Duration	Randomly assigned students	Randomly assigned teacher	Identical learning material	PBL procedure is stated clearly	Effect size
Kazemi and Ghoraishi (2012)	3 months	Yes	No	No	Yes	0.13
Ajai et al. (2013)	4 weeks	Yes	No	No	No	2.58
Imanieh et al. (2014)	4 months	Yes	No	No	Yes	1.43
Penjvini and Shahsawari (2013)	9 days	No	No	No	Yes	0.93
Westhues et al. (2014)	2 years	No	No	No	Yes	-0.61
Hendarwati et al. (2021)	1 semester	Yes	No	Yes	No	5.30
Firdaus and Herman (2017)	2 years	No	No	No	No	n/a
Aidoo et al. (2016)	3 months	Yes	No	No	No	1.96
Argaw et al. (2016)	1 week	No	No	No	No	0.74
Hendriana et al. (2018)	1 week	No	No	No	No	0.77
Amalia et al. (2017)	1 month	No	No	No	Yes	2.09

Source: Compiled by author

Table 2.5 states the duration, randomization procedure, treatment definition, and effect size (Cohen's d) of the relevant experimental studies. Randomized grouping is the first conundrum when designing experimental tests. Randomized grouping ensures that the students of treatment and control groups come from the same population so that any inherent differences can statistically be eliminated. Not all scholars correctly understand the concept of randomized grouping. For example, in two studies (Penjvini and Shahsawari, 2013; Westhues et al., 2014) which claimed that students are randomly assigned into PBL and control groups, significant differences can be observed in the mean value of their pretest scores. It should be noticed that even if there are no significant differences in pretest scores across PBL and control groups, this does not necessarily mean that they are randomized grouping. If students voluntarily joined the PBL program, as is the case in Westhues et al. (2014), they are different from non-PBL students in a way that may well be relevant, even if there is no difference in pretest scores.

The second challenge is the measurement of PBL. Section 2.1 of the present thesis demonstrated how flexible the PBL procedure can be. So, it is critical to clearly identify the PBL procedure used in the experiment. Half of the studies in Table 2.5 do not state the PBL procedure for the treatment group of their experiments. Even for those studies with an explanation of the PBL process, the reader has only a vague idea of it. For instance, Westhues et al. (2014) said the instructors received training in MacMaster, so we know that their PBL procedure belongs to the MacMaster camp but not more details. Without specific information on how much guidance is provided by the instructor, when the guidance is involved, and how many learners form each discussion group, the methods cannot be replicated by subsequent researchers. This challenge also indicates the advantage of breaking down PBL into individual elements in experiments, an empirical strategy which was suggested by Furtak et al. (2012).

It is also imperative to isolate the PBL treatment from other confounding factors. The two most common confounding factors are the teacher and the learning materials. In Table 2.5, most studies did not make sure that the teachers and learning materials are statistically identical between the treatment and control groups. Thus, we cannot know whether the observed difference in learning outcomes is driven by the PBL approach or the variances in teacher capability and the sutability of the learning materials. Such an empirical challenge is exacerbated by the fact that many PBL programs last for more than 1 week. In a two-year experiment like Westhues et al. (2014) and Firdaus and Herman (2017), it is all but impossible to insulate the experiment from confounding noises.

Not only PBL treatment but also the measurement of learning outcome faces empirical challenges. Previous studies (Matlen and Klahr, 2013; Rittle-Johnson, 2006) showed that PBL or its elements perform better in delayed testing than in immediate testing. Furthermore, PBL serves better for more intended learning goals than acquiring factual knowledge, including applying knowledge,[14] transferring knowledge, and critical thinking.[15] Some studies doubt

[14] Moallem (2019) (p. 108) claimed that "Rather than emphasizing the acquisition of knowledge and skills, PBL offers opportunities for students to apply knowledge and skills in the real world or an authentic context."

[15] Dabbagh (2019) (p. 153) stated that "PBL fosters the development of critical thinking skills such as problem-solving, analytic thinking, decision making, reasoning, argumentation, interpretation, synthesis, evaluation, collaboration, effective communication, and self-directed learning."

whether a testing score encompasses all the value of inquiry-based learning. Liou (2021) documented that although inquiry-based learning has negative direct effects on learning outcome, its overall effects are positive and greater than for direct instruction because of the indirect effect through learners' enjoyment and self-efficacy toward science. Rehmat and Hartley (2020) and Seibert (2021) argue that PBL can foster learners' critical thinking.

The last challenge of the empirical strategy is about the mediating mechanism, particularly cognitive load. As the discussions in Sections 2.2 and 2.3 indicate, it is likely that the empirical results based purely on learning outcomes cannot distinguish between certain different propositions. Including the mediation factor, cognitive load, into the empirical design can help since it is where traditional CLT postulates differently from social constructivism and modified CLTs. The potential drawback of cognitive load, that it can be only measured by subjective questionnaires, may hinder most PBL research from examining it. However, its value in distinguishing between competing theories has encouraged several studies (Kyun et al., 2013; Matlen and Klahr, 2013; Schmeck et al., 2015) to investigate its importance.

2.4.4 Summary of Section 2.4

This section reviews three categories of empirical strategies in PBL-related studies. Among these three categories, the pre-post non-experimental tests provide the least empirical value to the ongoing PBL-efficacy debates. Without a control group, it is difficult to exclude confounding factors from the findings of such tests. The second category of empirical strategy, large-sample-based association analyses, although having advantages such as fewer sample biases, data objectivity, and greater statistical power, still suffers from pitfalls including ambiguous causality inference and problems in its perception-induced measures.

Experimental or pseudo-experimental tests are the third category of empirical strategy reviewed in the current section. But such tests required careful controlling for any confounding factors through randomized grouping, clearly identified treatment factors, and strict control of confounding factors. The measurements of PBL and learning outcomes are also vital empirical issues in experimental design. The extant literature suggests the benefit of disaggregating PBL into single-factor elements and evaluating long-term learning outcomes. Testing the mediation effect of cognitive load is also recommended for distinguishing the underlying theoretical propositions. The discussions in the current section navigate the research design in Chapter 4.

2.5 Educational environment and PBL in China

The previous cross-country association analyses (see Section 2.3.1.1) indicate that the cultural factor could be an influential missing variable in learning outcomes; Choon-Eng Gwee (2008) and Frambach et al. (2012) argued that PBL efficacy varies across cultural contexts. The present thesis responds to the suggestion of Hung, Dolmans, et al. (2019) and investigates PBL efficacy in the context of non-western culture. Therefore, the current section briefly reviews the educational environment and PBL in China, especially for China's primary and secondary education, investigated by the present thesis.

2.5.1 China's top-down pedagogical reform since the 2000s

Before 2000, China's primary and secondary education was didactic and teacher-centered with the only purpose being to transfer standard knowledge to students (Paine, 1992). Officially, China's educational administration has launched a top-down pedagogical reform since the 2000s (Education, 2001, 2002, 2011). Such top-down pedagogical reform was aimed at improving teaching quality in China's primary and secondary classrooms. Influenced by the global trend of pedagogical method, China's new curriculum standards were pivoting toward a more student-centered and hands-off teaching approach which is in line with PBL principles (S. Gao et al., 2018; Guan and Meng, 2007; Ryan et al., 2009).

Although cases from other countries showed that top-down pedagogical reform may not be effective because of cultural factors (Du and Chaaban, 2020; Schweisfurth, 2013), some evidence suggests that China's reform achieve its desired effects. OECD (2011), using Shanghai as an exemplar, claimed that China's pedagogical reform "calls for an increase in the time allocated to student activities in classes relative to teachers' lecturing. This has caused a fundamental change in the perception of a good class, which was once typified by good teaching, with well-designed presentations by the teachers" (p. 34). Tan (2012) and Tan (2016) argued that the pedagogical reform in China has transformed school teaching practice into a model which is more student-centered and conducive to critical thinking. Sargent (2015) stated that despite the initially pessimistic attitude held by teachers, they eventually accepted the new pedagogical method by observing the positive effects. S. Gao et al. (2018) followed the TIMSS data standard to conduct association analyses for China's Inner Mongolia area and found that the hybrid

approach of lecture-based and inquiry-based learning methods is related to the best learning performance.

However, another group of researchers doubts whether the top-down pedagogical reform in China has really altered teaching methods in actuality. Dello-Iacovo (2009) criticizes that the reform brought about negligible changes in nationwide pedagogical practice, due to the conflict between the student-centered teaching method and China's examination-orientation system. Based on case studies of two chemistry classrooms in China, S. Gao and Wang (2014) concluded that Chinese teachers were reluctant to adopt the inquiry-based learning method promoted by pedagogical reform. Wang and Buck (2015) also suggested that the most important goal of teachers in China's primary and secondary schools is to prepare students to pass exams. Mostafa et al. (2018) ranked the index of implication frequency of inquiry-based science teaching methods in 56 OECD countries and found China with the sixth lowest score (p. 20, Figure 3.1). You (2019) attributed the difficulty of pushing top-down pedagogical reform in China to China's traditional educational system which is not in favor of the student-centered or PBL approach. The following section explores the historical roots for this in China's educational environment.

2.5.2 From Keju to Gaokao, China's historical root of the anti-PBL educational system

Ancient China has been well known for one meritocratic institution – the imperial examinations also named Keju – which was initially established in the late 6th century to screen and select civil servants for the government (Russell and Linsky, 2020). T. Chen et al. (2020) summarized three features of Keju: openness, an absence of corruption, and extreme competitiveness (p. 2035). Passing the Keju examinations meant immediate access to the higher strata of Chinese society and a significant increase in a person's economic and political status. Keju is therefore regarded as a way to foster social mobility and bolster the meritocratic political regime of ancient China (Stasavage, 2020). Education in ancient China was therefore Keju-exam-oriented.

Due to a series of defeats in foreign wars, the Qing Empire abolished the Keju system at the beginning of the 20th century (Bai, 2019). After that, China went through decades of war and turmoil until the founding of communist China. During the Mao era, the rules for social

advancement in communist China leaned significantly towards the 'Red' classes.[16] But as communist China opened up and downplayed the role of ideology in social institutions after 1978, China began to revert to its historical root of a meritocratic social system (Vickers and Xiaodong, 2017). A standardized college entrance examination, also named Gaokao, is one of the important mechanisms that inherited the characteristics of Keju and support a modern meritocratic social system in China (Liu, 2016).

Although Gaokao is not a civil service examination, it is comparable to Keju in terms of fairness and the importance of determining social class promotion. Like Keju, Gaokao strictly follows the anonymous marking rule and there is severe punishment for examination fraud. By participating in Gaokao, students compete for affordable but ranked tertiary educational resources[17] which largely determine their likelihood of becoming civil servants or employees in large state-owned enterprises after graduation; both are lucrative careers in China where the government heavily intervenes in the economy (Liu, 2013). To ascend to a higher social class, one of the main purposes of Chinese students' primary and secondary education has been to achieve better grades in Gaokao. In another word, China's modern primary and secondary education is Gaokao-exam-oriented.

Gaokao-exam-oriented primary and secondary education reduces the demands for learning higher-order knowledge such as knowledge application in reality. It is difficult to objectively evaluate the ability to apply knowledge in standardized tests such as Gaokao. An examination with standardized answers is also incompatible with social constructivism's tendency to allow for more individual innovation (You, 2019). Gaokao's function of encouraging social mobility allows it to sacrifice examining higher-order knowledge for fairness. The testing content of Gaokao, similar to Keju, focuses on a narrower range of basic knowledge (Muthanna and Sang,

[16] 'Red' classes are referred to as workers, peasants, and those from the political elite family who joined the communist party or army before 1949 (Li and Walder, 2001).

[17] Public universities are ranked in four tiers according to the educational resources they receive. Tier one is the "world-class" or Project-985 universities. Tier two is the top-100 or Project-211 universities. Tier three is comprehensive universities. Tier four is vocational and technical institutions. Although the quality of education at these public universities varies widely, they are all equally affordable to the majority population in China (Liu, 2013).

2016). It magnifies the advantage of knowledge retention in primary and secondary education and favors the traditional didactic teaching methods.

Besides the Gaokao-exam orientation system, other social factors may also make China's primary and secondary education less favorable to using PBL. For example, China's lower-value-added manufacturing economy used to demand a labor force with strict obedience rather than innovation (Wei et al., 2017). Meanwhile, China's strengthening autocratic regime is cautious about uncontrolled independent thoughts and critical thinking (Deudney and Ikenberry, 2009; Xue, 2021). All these political and economic factors, together with China's Gaokao-exam-oriented primary and secondary education environment, could undermine the value of PBL and hinder people to develop PBL-associated skills.

2.5.3 A tale of two educational paths in transitioning China

Besides the convention education path targeting Keju or Gaokao as the destination, there has been another education path in transitioning China, which targets studying abroad. Investing in education aiming to study abroad in developed countries used to be the dominant path in early 20th century China. Chinese students who graduated from universities in Europe and the United States occupied prominent political, economic, military, and cultural positions in China at that time. One student of the social constructivism forerunner John Dewey, Hu Shih, even led China's New Culture Movement (from the 1910s to 1920s) and reshaped the thoughts of the whole Chinese society (Grange, 2004). The investment return from the new education path in China was greater until the foundation of communist China.

The education path aiming to study abroad emerged again in China after its economic and opening-up reform in 1978. Even though the overall cost of learning abroad, both financial and non-financial, is much higher compared to the traditional Gaokao track, China has been the largest resource of international students in the world (UNESCO, 2019). According to the statistics of China's Ministry of Education (Education, 2020), 703,500 Chinese students were studying overseas in 2019, a 6.25% year-on-year growth. From 1978 to 2019, 6,560,600 Chinese chose the education path of studying abroad. The destinations of studying abroad are mainly Europe and US, which have, to a certain extent, implemented the PBL method for decades. The fever of Chinese parents sending children to study abroad could be explained by the attractive economic and social return (Cebolla-Boado et al., 2018; Y. Chen et al., 2021; Guo et al., 2019).

The two education paths in transitioning China result in different education demands (W. Zhang and Bray, 2017). The second path targeting learning abroad requires more critical thinking and innovative capability, which favors PBL. Those students who studied abroad after 1978 not only learned knowledge but also learn knowledge with PBL methods. Therefore, their overseas experiences changed their family-level cultural and educational concepts. The evidence from Malaysia showed that students who have had a PBL-like learning experience before are more likely to benefit from PBL (Jabarullah and Hussain, 2019). Similarly, the family-level cultural variance in China could also affect the efficacy of PBL.

2.5.4 Summary of Section 2.5

Previous studies suggest that PBL efficacy varies across cultural contexts. In the context of China, although there has been a top-down pedagogical reform toward PBL principles of learning since the 2000s, the effects of the reform were doubted. Existing research argues that the Gaokao-examination-oriented system inhibits the real implication of PBL principles in primary and secondary schools in China. China's Gaokao-oriented system can historically date back to ancient China's meritocratic institution named Keju. Due to Gaokao's function of encouraging social mobility, together with other social and economic factors in China, education sacrifices the goal of teaching higher-order knowledge such as critical thinking and innovation. However, another educational path of learning abroad exists in transitioning China, which is more favorable to the PBL approach. The choice between the two education paths in Chinese society thus contributes to the cultural variance at the family level, which in turn could affect the PBL efficacy based on evidence from other countries.

3 Research questions and hypotheses

With the review of previous studies on PBL in Chapter 2, two research gaps in this field have emerged. First, there is relatively little research on the effectiveness of PBL in non-Western cultural contexts. From a theoretical view, while traditional CLT does not propose a model incorporating cultural factors, constructivism theory claims that cultural factors could influence PBL's impacts on learning outcomes (see Section 2.2). The competing theories offer divergent predictions concerning the efficacy of PBL across different countries. From a practical perspective, PBL is an increasingly popular pedagogical approach worldwide. The propriety of disseminating the PBL approach globally requires validation through further experimental evidence, especially in light of the opposing inference that can be drawn from existing PISA-based association analysis (see Section 2.3). Thus, it is valuable — from both theoretical and practical standpoints — to raise research questions about PBL in non-Western cultural contexts.

Second, there is a need for a deeper understanding of the contextual factors that may influence the efficacy of PBL in general. Traditional CLT suggests that only two factors — students' prior knowledge and the complexity of learning tasks — will affect PBL efficacy, while constructivism and modified CLTs predict that there may be additional factors (see Section 2.3). Examining contextual factors outside the purview of traditional CLT can contribute to a more nuanced understanding of the relative veracity of different underlying theories. Examination of contextual factors can also provide educational practitioners with a more informed understanding of when PBL may be more useful in practice.

The two research gaps identified above, which have also been emphasized as areas of inquiry by systematic review studies such as Hung, Dolmans, et al. (2019), motivated the research questions of the present thesis. In particular, Research Questions 1 and 2 were intended to address the first gap by examining PBL in the non-western context of China. Research Questions 3 and 4 were designed to narrow the second gap by exploring the impact of three additional contextual factors in addition to those proposed by traditional CLT. The following sections of this chapter will outline the formulation of these research questions and the proposed testable hypotheses.

3.1 PBL efficacy in a non-Western country

PBL opponents, such as traditional CLT, typically do not anticipate variance in the effectiveness of PBL across countries (Sweller et al., 2019; Sweller, 2020). According to the framework of traditional CLT, as depicted in Figure 2.3, cognitive load is the sole influencing factor. To the best of my knowledge, there is no research within the scope of traditional CLT or modified CLTs that links the efficacy of PBL to cultural variations in short-term memory.[18] Therefore, we cannot infer from the framework of traditional CLT or modified CLTs that PBL efficacy will differ between Western and non-Western cultural contexts.

By contrast, constructivism theory places a strong emphasis on the role of social factors in the learning process from its very foundations (Schmidt et al., 2019). According to constructivism theory, knowledge must be situated within its proper context in order to be effectively conveyed. Furthermore, the context in which learning is situated is not only the immediate environment of the individual, but also the broader social-cultural context (Hung, Moallem, et al., 2019). PBL advocates such as Jonassen and Hung (2015) also argue that PBL models should be tailored to specific contexts, with the understanding that the same PBL approach may yield varying results in different cultural environments. As there is a discrepancy between the predictions of CLT and constructivism theory, conducting research on PBL efficacy in a non-Western cultural context and comparing the results to findings from Western countries thus can thus help to expand our understanding of PBL theory.

In practice, the adoption of PBL has been increasingly promoted by global educational administration agencies and has been implemented in a variety of disciplines and countries (Moallem, 2019). However, there is limited empirical evidence, with the exception of pilot studies without randomized control groups, to support the increasing adoption of PBL globally (L. Zhang et al., 2022). Association analyses of cross-country data (Aditomo and Klieme, 2020; e.g., Cairns and Areepattamannil, 2019; Forbes et al., 2020) have demonstrated significant variations in the pattern of guidance efficacy across countries, which suggests the potential for

[18] While there have been psychology studies (e.g., Alloway et al., 2017) examining the potential for variance in working memory capability across countries, these have not specifically addressed the relationship between cultural variations in short-term memory and PBL efficacy.

variations in the efficacy of PBL across different cultural environments, particularly between Western and non-Western countries. Nevertheless, the inherent shortcomings of association analysis techniques mean that they cannot fully substitute for the role of randomized controlled experimental research (see Section 2.4.2). Thus, it is important to provide experimental justification for the effectiveness of PBL in non-Western countries.

3.1.1 Formulation of RQ1 and RQ2

To formulate detailed research questions about the efficacy of PBL in a non-Western cultural context, I have chosen China as the specific research environment. China, with its unique social norms and the largest population[19] among non-Western countries, has a distinctive educational system that may have historically hindered the adoption of PBL (see Section 2.5.2). In the past two decades, the Chinese educational administration has implemented a top-down pedagogical reform that aligns with the principles of PBL by promoting a more student-centered and hands-off teaching approach (see Section 2.5.1). However, the effectiveness of this reform has been questioned. Specifying the research context within China can contribute to the growing but still insufficient research on PBL (S. Gao et al., 2018; e.g. Guan and Meng, 2007; Ryan et al., 2009) in this country.

An additional aspect to consider in formulating detailed research questions about the efficacy of PBL in a non-Western cultural context is selecting the appropriate education discipline. The education discipline, similar to the country, is a factor that constructivism theory posits as potentially impacting the efficacy of PBL, but which traditional CLT contends is inconsequential. Constructivist perspectives on education differentiate between disciplines that are 'well-structured,' such as mathematics, and those that are 'ill-structured,' such as literature, and argue that no single PBL approach is suitable for all disciplines (Jonassen and Hung, 2015; Moallem, 2019). On the other hand, traditional CLT classifies knowledge into primary and secondary biological categories, and does not believe that the discipline of study affects the ineffectiveness of PBL (Sweller et al., 2019; Sweller, 2021). In recent years, the disciplines of science, technology, engineering, and mathematics (STEM hereafter) are of particular interest due to the ongoing trend of incorporating PBL approaches in science education (L. Zhang et al.,

[19] Although India is projected to overtake China as the country with the largest population in mid-2023.

2022). The significance of science education as a discipline of study is also supported by the statistics obtained from the papers reviewed in Chapter 2.[20] Given the significance of the discipline of science, this thesis has chosen to focus on science education in its detailed research questions.

In formulating RQ1 and RQ2, this thesis aims not only to evaluate the overall effectiveness of a PBL approach, but also to examine the individual pedagogical elements that constitute PBL. As discussed in Section 2.1, PBL is composed of various individual elements. Three key features of PBL are 1) Optimal quantity of guidance, 2) Problem initiated before guidance, and 3) Small-group collaborative learning. There is furthermore a lack of theoretical and empirical evidence on the synergistic effect of using all of these elements together. The present thesis aims to gain a deeper understanding of PBL, by decomposing the effectiveness of PBL into its individual elements and testing the necessity of using all of these elements together.

In addition to decomposing the treatment variable, the present thesis also divides the learning outcome response variable into short-term and long-term portions. Previous empirical evidence suggests that PBL may have a particularly positive impact on long-term learning outcomes (see Sections 2.3 and 2.4.3). Accordingly, it is reasonable to assume that short-term and long-term learning outcomes may not always be the same. Short-term learning outcomes can be proxied by knowledge acquisition, which is also associated with cognitive load if the tenets of CLT hold (as cognitive load is posited to be the sole mediating variable affecting knowledge acquisition, see Figure 2.3). In contrast, indicators of long-term learning outcomes in previous empirical studies include students' enjoyment and self-efficacy. RQ1 focuses on how PBL affects short-term learning outcomes, while RQ2 investigates the potential for PBL to contribute to long-lasting learning effects. These research questions are formulated as follows:

Research Question 1: How do the key elements of PBL and the overall PBL pedagogical approach impact students' cognitive load and knowledge acquisition in the learning of science?

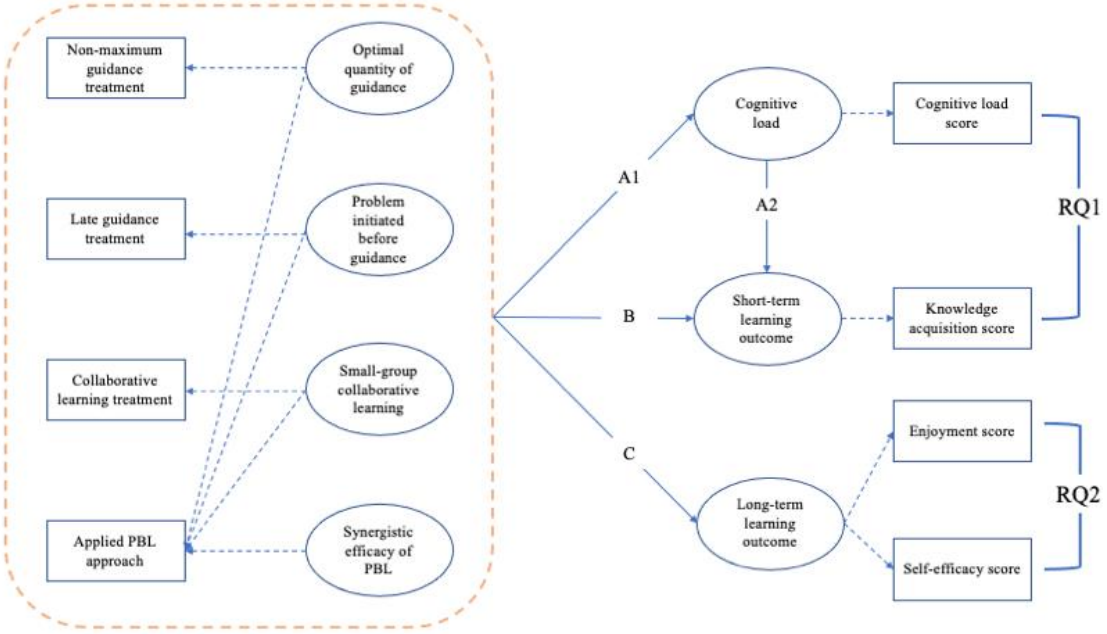
[20] Among the experimental studies presented in Chapter 2 (as listed in Figure 2.6, Figure 2.7, and Table 2.5), 37.8% were focused on science education, making it the most widely studied discipline in my reviewed sample. The second most popular discipline, which also belongs to the broad STEM discipline, was mathematics education with 32.4% of the studies focused on it. A detailed list of research studies organized by discipline can be found in Appendix 4.

Research Question 2: How do PBL's key elements and the overall PBL pedagogical approach affect students' enjoyment and self-efficacy in learning science?

3.1.2 Testable hypotheses of RQ1 and RQ2

The transformation of RQ1 and RQ2 into testable hypotheses involves the identification of causality paths and the selection of measurable variables. This process can be intuitively comprehended through the utilization of a diagrammatic representation, akin to that of a Structural Equation Modeling (SEM) diagram.[21] An illustration of such a SEM-like diagram is presented in Figure 3.1.

Figure 3.1: SEM-like diagram for RQ1 and RQ2



Source: Compiled by author

In Figure 3.1, the structural relationships among the latent and observable variables relevant to Research Questions 1 and 2 are depicted. The oval shapes in the diagram denote the latent variables, while the rectangles represent the corresponding measurable variables. The solid lines

[21] The SEM-like diagram provided here is a simplified version of a full SEM diagram, with the measurement error terms omitted for the purpose of concision. It can be considered as representing both the structural and partial measurement components of a traditional SEM diagram.

represent the causal relationships among the latent variables, and the dashed lines indicate the association between the latent variables and their corresponding measurable variables.

The left part of Figure 3.1 illustrates the treatment variables. It is noteworthy that there are four latent treatment variables: the three key feature elements previously discussed and the synergistic effect of PBL. The 'ideal' PBL approach posited by the constructivists can be considered as the optimal combination of these four latent treatments. 'Non-maximum guidance,'[22] 'Late guidance,' and 'Collaborative learning' are referred to as the three key PBL features, respectively. The PBL approach applied and observable in the experimental research reflects all three key PBL features and the synergistic effect.

The right part of Figure 3.1 depicts the response variables. RQ1 encompasses two latent variables, namely Cognitive load and Short-term learning outcome. The latent variable of 'Short-term learning outcome' is made manifest through the observable variable of 'Knowledge acquisition score,' which is derived from testing. Similarly, the latent variable of 'Cognitive load' is made manifest through the observable variable of 'Cognitive load score,' which is obtained through a questionnaire survey. RQ2 is concerned with the latent variable of 'Long-term learning outcome,' which can be quantified by two observable variables: 'Students' enjoyment score', obtained through a questionnaire survey, and 'students' self-efficacy', which is also obtained through a questionnaire survey.

The solid lines designated as 'A1' and 'A2' in Figure 3.1 represent the causality paths predicted by traditional CLT. In this sequence of pathways, cognitive load serves as a mediating variable. The solid lines designated as 'B' and 'C' represent the causality pathways predicted by the constructivism theory. The modified CLT, specifically ICLT, postulates the existence of both links 'A1->A2' and 'B' (see Section 2.2.4). Although constructivism theory and traditional CLT predict opposing signs of PBL efficacy, the empirical results may be affected by measurement errors for latent variables, particularly for the unobserved 'ideal' PBL approach. This highlights

[22] The first pedagogical element of PBL is the optimal level of guidance quantity. However, it is not possible to observe the optimal level of guidance directly. Instead, variations in non-maximum levels of guidance can be observed.

the significance of including cognitive load in the research questions and hypotheses, as it reflects the disparity between different theoretical frameworks.

The causality paths depicted in Figure 3.1 have been simplified. In theory, there should be solid lines emanating from each of the latent treatment variables, rather than a collective representation of PBL as a whole. As a result, the four observable variables on the left side and the four observable variables on the right side of Figure 3.1 would form 4x4 solid lines, leaving the figure cluttered and difficult to interpret. The causality predictions pertaining to each individual PBL feature element under various theoretical frameworks have been discussed in Section 2.3. Table 3.1 provides a summary of these predictions for convenient reference.

Table 3.1: Theoretical predictions regarding individual PBL feature elements

Treatment variable	Response variable	Theory	Prediction	Brief explanation
Non-maximum guidance	Knowledge acquisition	Traditional CLT and CCLT	-	The optimal level of guidance quantity is the maximum level
	Cognitive load	Traditional CLT and CCLT	+	Incomplete information increases extraneous load
	Knowledge acquisition	ICLT	+ or -	+ for expertise reversal effect and efficiency improvements in other phases; - for similar story of traditional CLT
	Cognitive load	ICLT	+ or -	Cognitive load could be increased in acquisition phase but reduced in other learning phases
	Knowledge acquisition	Constructivism	+ or -	+ with decreasing guidance quantity but - once the guidance quantity surpasses the optimal point; (or an inverted U-shaped relationship between the guidance quantity and learning efficiency)
	Cognitive load	Constructivism	?	Not concerning cognitive load
Late guidance	Knowledge acquisition	Traditional CLT and CCLT	-	Late guidance generates extraneous load so impairs learning efficacy
	Cognitive load	Traditional CLT and CCLT	+	Late guidance generates extraneous load
	Knowledge acquisition	ICLT	+ or -	+ for the theory of productive failure; - for the theory of inefficient failure increasing cognitive load
	Cognitive load	ICLT	+ or -	Cognitive load could be increased in acquisition phase but reduced in other learning phases
	Knowledge acquisition	Constructivism	+	An initial failed attempt to solve the problem can be conducive to consequent learning
	Cognitive load	Constructivism	?	Not concerning cognitive load

Treatment variable	Response variable	Theory	Prediction	Brief explanation
Collaborative learning	Knowledge acquisition	Traditional CLT and ICLT	?	Not concerning collaborative learning
	Cognitive load	Traditional CLT and ICLT	?	Not concerning collaborative learning
	Knowledge acquisition	CCLT	+ or -	Depends on the balance between benefits and costs of collaborative learning
	Cognitive load	CCLT	+ or -	+ because of communication among group members; - because of sharing cognitive resources among group members
	Knowledge acquisition	Constructivism	+	Interaction is beneficial for constructing knowledge; feedback and support also lead to positive emotional supports
	Cognitive load	Constructivism	?	Not concerning cognitive load

Source: Compiled by author

Table 3.1 provides a comprehensive overview of the predicted effects of the key elements of PBL on short-term knowledge acquisition and cognitive load. The table also includes brief explanations of the predictions under various theoretical frameworks. Sections 2.2 and 2.3 provide a more in-depth examination of the cognitive foundations and empirical evidence supporting these predictions. It is worth noting that the predictions about long-term learning outcome are not included in the table as only constructivism theory makes predictions about that. Table 3.1 also signifies the importance of including cognitive load in the research questions as it allows for the differentiation of underlying theoretical frameworks through empirical testing.

The thesis formulates testable hypotheses based on a theoretical framework akin to the ICLT. Specifically, it is posited that PBL and its individual elements, when compared to traditional didactic teaching methods, will result in enhanced knowledge acquisition scores, with the exception of an inverted U-shaped relationship between guidance level and short-term learning outcomes. Additionally, it is generally predicted that PBL and its individual elements will have a positive impact on enjoyment scores and self-efficacy scores, when compared to traditional didactic teaching methods. However, no predictions were made regarding the sign of cognitive load scores in this thesis. The predicted hypotheses related to RQ1 and RQ2 are summarized in Table 3.2. The symbol '+' denotes a prediction of positive effects resulting from the

implementation of the PBL approach or its individual component. Conversely, the symbol '-' denotes a prediction of negative effects. The symbol '?' indicates a lack of a clear prediction.

Table 3.2: Matrix of predicted hypotheses for RQ1 and RQ2

	Non-maiximum guidance treatment	**Late guidance treatment**	**Collaborative learning treatment**	**Applied PBL approach**
Cognitive load score	?	?	?	?
Knowledge acquisition score	+ for moderate guidance; - for minimal guidance	+	+	+
Enjoyment score	+	+	+	+
Self-efficacy score	+	+	+	+

Source: Compiled by author

In addition to the testable hypotheses documented in Table 3.2, the present thesis also examines the synergistic effect of PBL as an area of interest. However, it should be noted that this effect is not directly observable through a specific variable. The examination of this effect will be conducted through a three-way variance analysis or linear regression with three-way interaction terms, which will be discussed in further detail in Chapter 4.

3.2 Contextual factors of PBL efficacy

The second motivation for formulating the research questions of this thesis is to explore a wider range of contextual factors that may influence the effectiveness of PBL. Both CLT and constructivism acknowledge the existence of contextual factors that can impact the effectiveness of PBL. While traditional CLT suggests that certain contextual factors can mitigate the negative effects of PBL, constructivism posits that certain contextual factors can enhance the effectiveness of PBL. In comparison to traditional CLT, the theoretical frameworks of social constructivism and modified CLTs also allow for a broader range of contextual factors to be considered as influencing the efficacy of PBL (see Section 2.3). Investigating these contextual factors can aid in testing the underlying theories and provide a more nuanced understanding of the relative veracity of different theoretical frameworks.

Exploring contextual factors of PBL efficacy can also provide educational practitioners with a more informed understanding of when PBL may be more useful in practice. As discussed in

Section 2.1, the practical application of PBL has undergone evolution since its inception. Thus, identifying an optimal combination of PBL approaches and specific contextual factors can be beneficial for PBL practitioners. As suggested by Hung, Dolmans, et al. (2019), research should shift its focus to understanding the circumstances under which PBL is effective or not. This thesis adheres to this guidance by not only addressing the fundamental question of 'Does PBL work?' in a non-western cultural context, but also striving to answer the question of 'When and why does PBL work?' in general.

3.2.1 Formulation of RQ3 and RQ4

RQ3 and RQ4 are formulated based on different types of contextual factors that may influence the effectiveness of PBL. RQ3 specifically examines the impact of students' prior knowledge and learning-task complexity on the efficacy of the overall PBL approach. According to traditional CLT, the negative impact of PBL is less pronounced when learners have more prior knowledge or the learning task is less complex (O. Chen et al., 2017; Sweller et al., 2011). PBL advocates also agree that these factors can affect the efficacy of PBL, but from a benefit perspective, such as Richey and Nokes-Malach's (2013) argument that withholding instructional explanations fosters students' constructive cognitive activities. It is therefore important to examine the impact of these contextual factors on science learning in China, and RQ3 is formulated as follows:

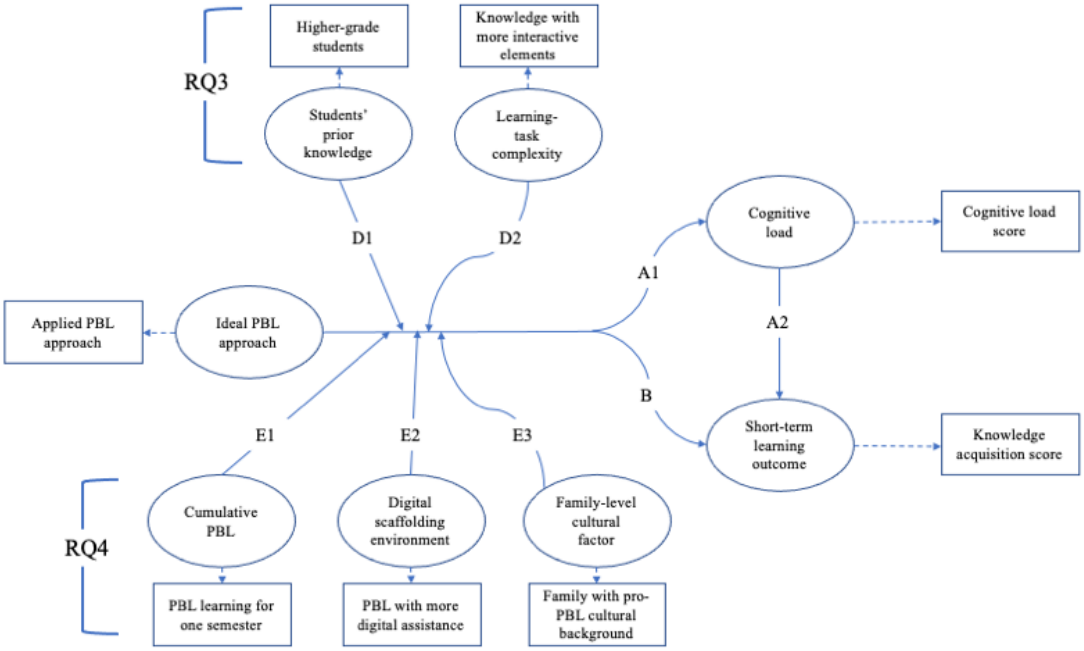
Research Question 3: Do students' prior knowledge and learning-task complexity influence PBL's efficacy on students' cognitive load and knowledge acquisition in learning science?

RQ4 focuses on investigating the impact of additional contextual factors on the efficacy of PBL. The theoretical frameworks of social constructivism and modified CLTs allow for a broader range of contextual factors to be considered as potentially influencing the effectiveness of PBL. While current empirical studies on the effects of PBL have primarily focused on factors within the scope of traditional CLT, this research aims to explore a wider range of potential influential factors, including those discussed in the literature, such as Moallem (2019). RQ4 is formulated as follows:

Research Question 4: Do other factors influence PBL's efficacy on students' cognitive load and knowledge acquisition in learning science?

In formulating RQ3 and RQ4, my focus is on the overall treatment of PBL rather than its individual elements, in order to make the research more appropriate for the classroom. RQ3 and RQ4 also limit the examination to short-term learning outcomes as response variables, as CLT does not account for the effects of contextual factors on long-term learning outcomes indicators. Despite this, cognitive load remains an important response variable in RQ3 and RQ4, as it enables us to distinguish between different underlying theories. Similar to RQ1 and RQ2, I employ a SEM-like diagram for RQ3 and RQ4, as depicted in Figure 3.2.

Figure 3.2: SEM-like diagram for RQ3 and RQ4



Source: Compiled by author

Figure 3.2 illustrates the representation of additional contextual factors that are considered in RQ4. This figure also illustrates the observable variables associated with these latent variables. Figure 3.2 serves as a guide for the examination of testable hypotheses pertaining to RQ3 and RQ4 in the following section.

3.2.2 Testable hypotheses of RQ3 and RQ4

In Figure 3.2, the latent variables are denoted by oval shapes, while the observable variables are represented by rectangles. The causal relationships between the variables are again marked by

solid lines, with dashed lines indicating the connection between latent variables and their corresponding observable variables. The causality paths indicated by A1, A2, and B pertain to the efficacy of PBL on learning outcomes (as previously discussed in Figure 3.1). The causality paths symbolized by D1, D2, E1, E2, and E3 pertain to the impact of contextual variables on PBL efficacy.

The solid lines D1 and D2 in Figure 3.2 represent the two contextual factors identified by traditional CLT and examined by RQ3 of this thesis. CLT postulates that the negative effects of PBL will be less pronounced if the learners have more prior knowledge or the learning task complexity is lower. These propositions, referred to as the Expertise reversal effect and Element interactivity effect, occur when the intrinsic load is small, allowing increased extraneous load to not deplete working memory. It is important to note that learning task complexity is not synonymous with task difficulty. As per previous research, this thesis uses higher grade students as a proxy for students with more prior knowledge and element interactivity as the observable variable for learning task complexity. This thesis predicts that, for higher grade students and learning tasks with fewer interactive elements, students are likely to achieve higher knowledge acquisition test scores.

The solid lines E1, E2, and E3 in Figure 3.2 signify three additional contextual factors that are examined by this thesis, beyond the two contextual factors considered in traditional CLT. E1 denotes the cumulative PBL effect, which suggests that the positive impacts of PBL are more pronounced for students who have more experience with PBL pedagogy. Previous studies such as Dolmans et al. (2016) have explored the impact of the learning environment, specifically the extent of PBL implementation (single course vs. curriculum-wide), on the effectiveness of PBL and found that a curriculum-wide implementation of PBL has a more positive impact on deep learning compared to implementation within a single course. To examine this cumulative PBL effect, this thesis employs prior experience of PBL for one semester as a measurable variable.

The solid line E2 represents the inclusion of digital scaffolding as a contextual variable in the examination of PBL's efficacy. Kim et al. (2018) utilize Bayesian meta-analysis to investigate the effectiveness of computer-based scaffolding in the context of PBL for STEM education. Their results indicate that computer-based scaffolding has a significant positive impact on improving learning efficiency. In line with the findings of Kim et al. (2018), this thesis predicts

that PBL supplemented with digital scaffolding techniques will result in better performance on knowledge acquisition tests, in comparison to PBL without such assistance.

The solid line E3 in Figure 3.2 includes the family-level cultural factor as a contextual variable that influences the efficacy of PBL. Forbes et al. (2020) specifically found that Korean students are less likely to engage in argumentation or draw conclusions, suggesting that cultural norms play a significant role in shaping pedagogical practices. Similarly, evidence from Malaysia has shown that students who have had previous PBL-like learning experiences are more likely to benefit from PBL instruction (Jabarullah and Hussain, 2019). In the case of China, the present thesis posits that the efficacy of PBL may be affected by family-level cultural variance, as families in China often have vastly different education demands and experiences due to their differing educational paths (see Section 2.5). As such, it is predicted that for families with a pro-PBL cultural background, the efficacy of PBL will be more pronounced.

The effects of the five contextual factors on acquisition have been analyzed and discussed in the preceding paragraphs of this section. With regard to the impacts on cognitive load, due to the conflicting evidence from previous studies (see Sections 2.3.1.2 and 2.3.2), this thesis does not make predictions for the effects of 'Higher-grade students' and 'Knowledge with more interactive elements.' However, this thesis posits that 'PBL for one semester,' 'PBL with more digital assistance,' and 'Family with pro-PBL cultural background' are likely to reduce cognitive load. Table 3.3 summarizes the predicted hypotheses related to RQ3 and RQ4. The symbol '+' denotes the presence of enhanced positive effects or mitigated negative effects resulting from the implementation of the PBL approach in the presence of the contextual variable. Conversely, the symbol '-' denotes the presence of mitigated positive effects or enhanced negative effects. The symbol '?' indicates a lack of a clear prediction or determination of effects.

Table 3.3: Theoretical predictions regarding the PBL context

Contextual variable	Response variable	Prediction
Higher-grade students	Cognitive load score	?
	Knowledge acquisition score	+
Knowledge with more interactive elements	Cognitive load score	?
	Knowledge acquisition score	-

Contextual variable	Response variable	Prediction
PBL for one semester	Cognitive load score	-
	Knowledge acquisition score	+
PBL with more digital assistance	Cognitive load score	-
	Knowledge acquisition score	+
Family with pro-PBL cultural background	Cognitive load score	-
	Knowledge acquisition score	+

Source: Compiled by author

3.3 Summary of Chapter 3

This chapter begins by summarizing the research gaps that emerged from the literature review in Chapter 2. Motivated by these gaps, the section identifies two research opportunities: 1) the examination of the effectiveness of PBL in a non-Western cultural context and 2) the investigation of the factors that contribute to the success or failure of PBL in general.

To address these two research opportunities, the chapter formulates four detailed research questions by specifying the country and discipline as the research background. These research questions are also formulated by specifying the latent treatment and response variables. To convert these research questions into testable empirical hypotheses, the chapter identifies the causality links to be tested and the measurable variables or treatments for the latent variables. Two SEM-like diagrams (Figures 3.1 and 3.2) are employed to depict the relationship of the various variables to the research questions. This chapter concludes by summarizing the testable empirical hypotheses in Tables 3.2 and 3.3. These hypotheses lay the foundation for the research design in Chapter 4.

4 Research design

To address the four research questions and related testable hypotheses presented in Chapter 3, the present thesis utilized a research design comprising six experiments. Table 4.1 provides a comprehensive overview of the various experiments conducted. This table also indicates the research questions being addressed in each experiment, the focus of the study, the school where the experiment was conducted, the grade level of the students involved, the time period, and the learning task of the experiment. The experiments were conducted at two different schools: a public middle school in Naning (Luzhou Huojing Zhan School; LHZ School hereafter) and a private middle school in Wenzhou (Ruian Zijing Shuyuan School; RZS School hereafter). The experiments took place between 2016 and 2018 and involved students from grades 8 and 9.

Table 4.1: List of six experiments

Experiment	Research Questions	Focus	School	Grade	Time	Learning task
1	RQ1, RQ2	Effect of PBL pedagogical elements and overall PBL approach	LHZ School	Grade 8	Mar 2016	Newton's laws of motion
2	RQ1, RQ2	Interaction effect of three PBL pedagogical elements	LHZ School	Grade 8	Feb 2017	Newton's laws of motion
3	RQ3	Contextual effects of students' prior knowledge	RZS School	Grade 8, Grade 9	Sep 2016	Newton's laws of motion
4	RQ3	Contextual effects of complexity of learning knowledge	RZS School	Grade 8	Sep 2017	Newton's laws of motion and conservation of energy
5	RQ4	Contextual effects of previous PBL experience and digital scaffolding techniques	LHZ School	Grade 8	Dec 2018	Newton's laws of motion
6	RQ4	Contextual effect of pro-PBL family cultural effects	RZS School	Grade 8	Sep 2018	Newton's laws of motion

Source: Compiled by author

Two types of organization structure could be chosen for this chapter. The first option is to present the research design experiment by experiment, while the second option is to present the

design by elements and group experiments under each element. Given the fact that the six experiments share many common issues and are not entirely distinct, this chapter has chosen the latter method of organization.

In particular, the organization of this chapter is as follows. Section 4.1 introduces the randomization strategy utilized in this thesis to mitigate potential imbalances between treatment and control groups through conventional sampling techniques. Section 4.2 outlines the measurement constructs of the variables. Section 4.3 provides information on the experimental designs and statistical analysis models employed in this thesis. Finally, Section 4.4 outlines the experimental process and provides a brief statistical description of the sample.

4.1 Stratified random sampling

A crucial prerequisite issue in the design of my thesis research is the establishment of randomized controlled groups. Simple randomization techniques may not account for systematic differences between treatment and control groups, particularly with regard to pre-treatment testing scores or other confounding factors that may impact student learning outcomes. Such systematic disparities have been observed in Penjvini and Shahsawari (2013) and Westhues et al. (2014), as discussed in Section 2.4.3. In order to mitigate this issue, all six experiments in my thesis utilize a stratified random sampling strategy based on students' most recent expected learning outcomes. This approach has been advocated by previous research in the field of sociology, public health, and education (Cochran, 2011; L. Cohen et al., 2002).

The sampling process employed in my thesis involves two steps. First, I estimate students' expected learning outcomes using a linear mixed-effects model (LMM hereafter). For each experiment, all students who are eligible to participate in the study are included in the regression. The dependent variable is students' final testing scores in the most recent semester, while predictors include student age, gender, and family socio-economic status index, which have been shown in prior research (see Section 2.3.1.1) to influence student learning outcomes. The random effect of student's previous classroom is also considered in the regression model. By using fitted

values as opposed to actual learning outcomes, I aim to control for potential measurement errors.[23] The regression formula of LMM can be represented as:

$$RPF_{ij} = (\gamma_{00} + \zeta_{0i}) + \beta_1 \cdot Age_{ij} + \beta_2 \cdot Gender_{ij} + \beta_3 \cdot SESI_{ij} + \epsilon_{ij} \qquad (4.1)$$

where the subscripts i and j denote student j in class i. Thus ζ_{0i} captures the clustering effects of students' previous class.[24] RPF (Recent physics finals) is the student's actual physics final testing score in the latest semester. Age, $Gender$, and $SESI$ are student's age, gender, and family socio-economic status index respectively. The details of variable measurement will be explained in Section 4.2.

In the second step, students are randomly assigned to experimental groups according to strata formed on the basis of expected learning outcomes derived in step one. I rank students' expected learning outcomes from highest to lowest, and then divide them into strata. The size of each stratum depends on the number of treatment and control groups[25] required by the experiment. For example, if there is only one control group and one treatment group, each stratum would contain two students. If there are seven experimental groups and no control group, as in my first experiment, each stratum would contain seven students. Within each stratum, student assignment is randomized. This stratified random sampling approach serves to minimize disparities between treatment and control groups. The process of stratified random sampling used in my thesis is depicted in Figure 4.1.

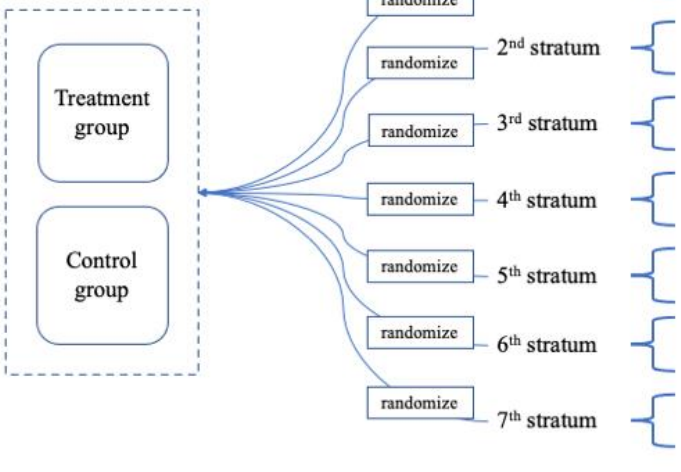
---

[23] This method is conceptually similar to propensity score techniques, which also aim to control for confounding factors, although logistic regression is typically used in those cases. Thoemmes and Kim (2011) offer a systematic review of the utilization of the propensity score method in the field of social sciences.

[24] Equation 4.1 represents a LMM, which differs from the conventional approach of incorporating i-1 indicator variables in a regression analysis. This model alters the expected variance-covariance matrix of ϵ_{ij} by incorporating the standard deviation of ζ_{0i}. In contrast, incorporating i-1 class indicators preserves the expected covariance of ϵ_{ij} as zero. LMM provides a larger testing power compared to the conventional approach, as it allows for an increased number of degrees of freedom (Galecki et al., 2013). LMM has been commonly used by PBL association analysis studies (Areepattamannil, 2012; Cairns and Areepattamannil, 2019; Kaya and Rice, 2010; Lau and Lam, 2017; L. Zhang and Li, 2019) with an alternative name: hierarchical linear model.

[25] In Experiment 6 and the first part of Experiment 5, a revised stratified random sampling methodology was utilized due to the unbalanced distribution of sample sizes with respect to contextual factors. Further details can be found in Section 4.4.

Figure 4.1: The process of stratified random sampling



Student	Predicted learning outcome	Rank
A	93	1
B	92.3	2
C	91.4	3
D	87.4	4
E	86.8	5
F	83.7	6
G	77	7
H	75.8	8
I	75.6	9
J	74.5	10
K	69.7	11
L	69.1	12
M	64.3	13
N	60.7	14

Source: Compiled by author

As depicted in Figure 4.1, the sample population was ranked according to their predicted learning outcome, with the highest predicted outcome at the top and the lowest at the bottom. Two experimental groups, one treatment and one control group, were established within the sample population, with each stratum containing two students. The assignment of students to treatment and control groups within each stratum was conducted randomly. This method of stratified random sampling ensures that the predicted learning outcomes were matched across the experimental groups.

4.2 Measurement constructs of empirical variables

Sections 3.1.2 and 3.2.2 outlined the empirical variables in the testable hypotheses of this thesis. In order to provide a comprehensive understanding of the research conducted, this section will delve into the measurement constructs of these empirical variables. Table 4.2 provides a summary of the definitions and constructions of the variables used in this thesis.

Table 4.2: Variable table

Variable category	Variable name	Variable definition and construct
Treatment	MaxGuid	Maximum guidance treatment. Equal to 1 if the instructor provided a pre-recorded video lecture that contained a maximum level of detail, and 0 otherwise.
	ModGuid	Moderate guidance treatment. Equal to 1 if the instructor provided a pre-recorded video lecture that contained a moderate level of detail, and 0 otherwise.
	MinGuid	Minimal guidance treatment. Equal to 1 if the instructor provided a pre-recorded video lecture that contained a minimal level of detail, and 0 otherwise.
	LateGuid	Late guidance treatment. Equal to 1 if the first video, in which the instructor poses daily-life questions related to Newton's law, was played to students prior to the second video, in which the instructor provides a lecture on the same topic, and 0 otherwise.
	Collective	Collective learning treatment. Equal to 1 if students were prompted to engage in small group discussions (consisting of three students per group) while viewing the lecture video, and 0 otherwise.
	Didactic	Didactic teaching treatment. Equal to 1 if students received maximum guidance without late guidance or collective leaning, and 0 otherwise.
	PBL	Applied PBL treatment. Students received moderate guidance, late guidance, and collective leaning. Equal to 1 if Applied PBL treatment was conducted and 0 otherwise.
	Placebo	Placebo treatment. Students were presented with a didactic teaching method. However, they were informed that they were participating in an experimental trial of a new learning approach. Equal to 1 if Placebo treatment was conducted and 0 otherwise.
Response	IScore	Students' immediate test scores, which are acquired from an assessment of the learning task immediately after the completion of the learning activity, with a possible score range of 0 to 100.
	DScore	Students' delayed test scores, which are obtained from an evaluation of the learning task conducted 7 days after the completion of the learning activity. The scores can range from 0 to 100.
	CogLoad	The cognitive load score of the students, which is a latent variable determined through the application of CFA on items CL1 through CL8.
	Enjoy	The enjoyment score of the students, which is a latent variable determined through the application of CFA on items E1 through E5.
	SelfEff	The self-efficacy score of the students, which is a latent variable determined through the application of CFA on items S1 through S5.
Contextual	Hgrade	Higher grade student, a binary variable with a value of 1 indicating the student is in 9th grade and a value of 0 indicating the student is in 8th grade.
	Complex	Learning task complexity, which is a binary indicator, where a value of 1 signifies that the task pertains to both Newton's law and conservation of energy, and a value of 0 denotes that the task pertains solely to Newton's law.
	CumuPBL	Students' prior experience with PBL, which is a binary indicator, where a value of 1 indicates that the participating student had completed at least one semester of PBL prior to the experiment, and a value of 0 otherwise.
	Digital	Digital assistance, a binary variable, with a value of 1 indicating the use of interactive simulation software during the instruction of Newton's law or conservation of energy, and a value of 0 indicating otherwise.
	ProPBL	Pro-PBL family cultural environment, which is a binary variable, with a value of 1 indicating that the student's father or mother had prior overseas education experience, and a value of 0 indicating otherwise.
Covariate	Gender	Student's gender, which is a binary variable, where a value of 1 indicates the student is male and a value of 0 indicates the student is female.

Variable category	Variable name	Variable definition and construct
	Age	Student's age at the time of participating in the experiment, measured in years.
	SESI	The socio-economic status of the student's family, as determined by a weighted average of responses to a 10-question survey, with a possible range of values from 0 to 10.
	RPF	The student's score on their most recent physics semester final exam, with a possible range of values from 0 to 100.

Source: Compiled by author

The names of the variables listed in Table 4.2 will be utilized throughout this thesis. Table 4.2 categorizes the empirical variables of this thesis into four categories: treatment, response, covariate, and contextual variables, which will be explained in more detail in the following sections.

4.2.1 Treatment variables

As previously discussed in Section 2.4.3, one of the key empirical challenges in this thesis is the isolation of the treatment effect. The treatment variables in this research pertain to the PBL approach and its individual elements. As indicated in Table 2.5, a majority of the studies did not effectively ensure that the instructors and learning materials were appropriately matched for the treatment and control groups. With the presence of excessive noise in treatment variables, the observed efficacy of the treatment may not be attributed solely to the PBL pedagogical approach, but may also be influenced by confounding factors such as teaching duration, instructor, or materials.

To effectively isolate the treatment effect, it is crucial to ensure that the teaching environment is consistent across the treatment and control groups. The methodology employed in this thesis is similar to that of Ashman (2022), where pre-recorded teaching videos were presented to students. Specifically, four different types of teaching videos were prepared, including those with maximum guidance, moderate guidance, minimal guidance, and those that introduce initial problems. The corresponding example slide descriptions for these video elements can be found in Appendices 11, 12, 13, and 14, respectively. The baseline learning task in the experiments pertains to the study of Newton's laws of motion in the physics discipline. For each experiment, the teaching videos were recorded by a single instructor. The video elements were then combined in different sequences to form various teaching treatments.

Experiment 1 further included a placebo control group, to rule out the potential psychological impacts on learning not related to PBL treatments. This approach, commonly used in medical research, has also been applied in education studies to control for placebo effects, where students may exhibit increased mental effort due to being aware of a special learning treatment (Fraenkel et al., 2012, p. 280). Lipsey and Wilson (1993) found that the average effect size of previous psychology and education studies may be inflated by 40% without the inclusion of a placebo control group. This method has been employed in several recent educational experimental studies, such as the use of a general learning strategy training as a placebo group to assess the effectiveness of multimedia instruction (Scheiter et al., 2015) and the use of Turkish reading activities as a placebo group to evaluate the effectiveness of inquiry-based mathematics learning (Karademir and Akman, 2019). In the placebo group for Experiment 1, students were shown a recorded video with explicit teaching, which was the same as the didactic teaching group, but were told that they were participating in an innovative learning experience.

Table 4.3 provides the detailed processes of the treatment variables. As demonstrated in Table 4.3, each group received pedagogical guidance through pre-recorded instructional videos, which were presented in three variations based on the level of direction offered. To mitigate the impact of varying learning durations, the duration of instruction and the class size[26] were standardized across all groups, thereby controlling for any discrepancies that may arise as a result of differences among teachers.

Table 4.3: Treatment detailed processes

Treatment	Duration	Class size	Detailed process
Didactic teaching	30 minutes	50 students	a) Students watched the previously recorded video with a maximum level of detail b) Students watched the previously recorded video with daily-life questions related to learning subject

[26] In the implementation of the experiments, three students from the total of 1,450 participants failed to attend the delayed tests and were thus excluded from the final sample. This resulted in slight variations in class size across the groups, which are no longer strictly equal to 50. Further details on this matter can be found in Section 4.4.

Treatment	Duration	Class size	Detailed process
Didactic teaching (placebo)	30 minutes	50 students	a) Students watched the previously recorded video with a maximum level of detail, but were told that they were attending an innovative learning experience b) Students watched the previously recorded video with daily-life questions related to learning subject
Least guidance	30 minutes	50 students	a) Students watched the previously recorded video with a minimal level of detail b) Students watched the previously recorded video with daily-life questions related to learning subject
Moderate guidance	30 minutes	50 students	a) Students watched the previously recorded video with a moderate level of detail b) Students watched the previously recorded video with daily-life questions related to learning subject
Late guidance	30 minutes	50 students	a) Students watched the previously recorded video with daily-life questions related to learning subject b) Students watched the previously recorded video with a maximum level of detail
Collaborate learning	30 minutes	50 students	a) Students watched the previously recorded video with a maximum level of detail b) Students discuss in 3-person groups c) Students watched the previously recorded video with daily-life questions related to learning subject
Applied PBL	30 minutes	50 students	a) Students watched the previously recorded video with daily-life questions related to learning subject b) Students watched the previously recorded video with a moderate level of detail c) Students discuss in 3-person groups

Source: Compiled by author

As demonstrated in Table 4.3, the students in the various experimental groups only viewed pre-recorded videos created by the same instructor. The only source of variation between the experimental groups is the choice and order of videos, which I designed intentionally. The treatment group represents the intended treatment factor. The Applied PBL treatment consists of a combination of moderate guidance, late guidance, and collaborative learning, as outlined in Section 2.1 and supported by the inverse relationship between guidance level and learning outcomes documented in Section 2.3.

4.2.2 Response variables

In this thesis, the construct of acquisition knowledge score was operationalized by administering two assessments on the learning content to the students. The first assessment was administered

immediately following the learning, while the second assessment was conducted after a 7-day interval. Samples of the assessment questions can be seen in Appendix 15. Previous research (Ashman et al., 2020; Chase and Klahr, 2017; Hsu et al., 2015; Matlen and Klahr, 2013; Steenhof et al., 2020) has suggested that immediate test scores may differ from delayed test scores in terms of PBL efficacy. As such, I did not convert the immediate and delayed test scores into a single factor using CFA. Instead, both scores were utilized as proxies for the response variable of acquisition knowledge score in the analyses of this thesis.

The construct of cognitive load has been widely discussed in the literature, with studies such as DeLeeuw and Mayer (2008), Leppink et al. (2013), Leppink et al. (2014), Ayres (2017), and D. Jiang and Kalyuga (2020) providing insight into its measurement. Typically, cognitive load is determined through a set of questions, which may be classified as either subject or object, and as either direct or indirect. The answers to these questions may take the form of numerical or pictorial scales. As noted by Ouwehand et al. (2022), Likert Rating Scales remain a preferred method for measuring cognitive load. Additionally, Schmeck et al. (2015) suggests that it is optimal to evaluate cognitive load immediately following a learning task. In this thesis, eight questions were designed to assess cognitive load, with each question offering four answer options ranging from ‘strongly agree’ to ‘strongly disagree.’ These questions can be found in Appendix 8.

In order to quantify Cognitive Load, the eight cognitive items were first converted into numerical values, with a score of 4 indicating the highest cognitive load and 1 indicating the lowest cognitive load. Previous studies, such as D. Jiang and Kalyuga (2020), have used CFA to generate intrinsic and extraneous cognitive load factors. However, this thesis does not distinguish between these two factors and instead focuses on the total cognitive load. Therefore, CFA was applied to the eight cognitive items to convert them into a single cognitive load variable $CogLoad$.

Similarly, the response variables of students’ Enjoyment and Self-efficacy were also measured through Likert Rating Scale questions. These questions were based on those used in the PISA 2006 and 2015 studies (OECD, 2009, 2017), with adjustments made to ensure they were appropriate for the field of physics. Both enjoyment and self-efficacy were assessed through five questions, which can be found in Appendices 9 and 10 respectively. After coding the students’

answers into numerical values, CFA was performed to extract a single factor for both students' enjoyment and self-efficacy, which were labeled as $Enjoy$ and $SelfEff$, respectively. The details of constructing CFA latent variable can be found in Appendix 7.

Table 4.4 displays the reliability assessment results for Cognitive Load, Enjoyment, and Self-efficacy. Reliability testing results for the indicators of Enjoyment and Self-efficacy are not present in Experiments 3 to 6, as the research questions RQ3 and RQ4 do not encompass these long-term learning outcome measures (see Section 3.2). Two metrics, Cronbach's Alpha (Cronbach, 1951) and McDonald's Omega (McDonald, 2013), were calculated based on the data collected in each experiment. McDonald's Omega was also employed in this study, as it has more relaxed assumptions regarding the latent variable model compared to Cronbach's Alpha (Dunn et al., 2014; Revelle and Zinbarg, 2009). The results of the reliability testing indicate that the Cronbach's Alpha and McDonald's Omega values are all above 0.8, suggesting a high level of reliability for the observed questionnaire items.

Table 4.4: Reliability testing results for CogLoad, Enjoyment, and Self-efficacy

Latent variable	Items	Experiment	Alpha	Omega
CogLoad	*CL1 - CL8*	1	0.96	0.95
		2	0.97	0.96
		3	0.97	0.95
		4	0.97	0.96
		5	0.96	0.94
		6	0.97	0.95
Enjoy	*E1 - E5*	1	0.86	0.82
		2	0.86	0.83
SelfEff	*S1 - S5*	1	0.86	0.82
		2	0.84	0.81

Source: Compiled by author

4.2.3 Covariate and contextual variables

In this thesis, four confounding covariates were taken into consideration: students' age, gender, recent Physical Science exam score, and socio-economic status, which are labeled as Age, $Gender$, RPF, and $SESI$ respectively (see Table 4.2 for detailed definitions). The inclusion of student socio-economic status in research is a common practice in studies based on PISA and TIMSS, and this factor has been demonstrated to have a positive impact on student learning outcomes (Areepattamannil, 2012; Cairns and Areepattamannil, 2019; Jerrim et al., 2020). The SESI construct in this study was modeled on a scoring system similar to the one used in PISA2016 (OECD, 2017), with higher scores indicating higher socio-economic status. The methodology used to construct the SESI, including the ten questions and their weightings, can be found in Appendix 6.

While the impact of gender on student learning outcomes is not always significant, some studies (e.g., Cairns and Areepattamannil, 2019) suggest that male students are more interested in science learning. Although age can theoretically influence a student's learning capability, there is limited variance in student age in this research, making it an optional covariate to control. Nevertheless, age was included in the model. Previous testing performance, as controlled in research such as Jerrim et al. (2020), was also controlled for. These confounding covariates were used for stratified random sampling, as explained in Section 4.1, and in the variance analysis and linear regression models, which will be discussed in Section 4.3.

Section 3.2.2 has identified five contextual variables that serve as the focus of this thesis: Higher-grade students, Knowledge with more interactive elements, PBL for one semester, PBL with more digital assistance, and Family with pro-PBL cultural background. In this section, each of these variables will be examined in greater detail.

Experiment 3 focuses on the contextual variable of higher-grade students (labeled as $Hgrade$ in Table 4.2). $Hgrade$ is measured using a binary variable, with a value of one assigned to 9th-grade students at RZS School in September 2016 and a value of zero assigned to 8th-grade students at the same school and time. According to the curriculum at RZS School, neither 8th-grade nor 9th-grade students had prior knowledge of Newton's law of motion. Therefore, the learning task was new to both groups. However, 9th-grade students had previously acquired

knowledge of basic concepts such as velocity and acceleration, which made them experts within the CLT framework.

Experiment 4 is focused on the contextual variable of knowledge that incorporates more interactive elements, referred to as $Complex$ in Table 4.2. $Complex$ is represented by a binary indicator, with a value of one indicating that the task pertains to both Newton's law and the conservation of energy, and a value of zero indicating that the task pertains solely to Newton's law. There is a close relationship between Newton's law of motion and the conservation of energy, but incorporating more interactive elements is necessary in order to fully grasp and connect these concepts. Therefore, students who learn both Newton's law of motion and the conservation of energy together can be considered as engaging in learning tasks with a higher level of interactivity within the framework of CLT.

In Experiment 5, two contextual variables are analyzed: $CumuPBL$ (representing prior PBL experience for one semester) and $Digital$ (representing PBL with an increased level of digital assistance). $CumuPBL$ is a binary variable that signifies the existence or absence of prior PBL experience. The LHZ School initiated a pilot teaching program in the Autumn semester of 2018, which was supervised by the Guangxi Education Department and was designed to reflect a PBL-style curriculum. During the time of Experiment 5 in December 2018, three classes of Grade-8 students had been participating in the pilot program for over three months. Consequently, students with prior PBL-like learning experience were designated as $CumuPBL = 1$, while those without such experience were designated as $CumuPBL = 0$.

$Digital$ is a binary variable that signifies the utilization of interactive simulation software. The software was developed by Beijing Rainer Software Technology Co., Ltd., a Chinese company specializing in providing virtual reality and interactive simulation teaching applications for primary and secondary education. A sample demonstration of the use of this software to teach Newton's law of motion can be found in Appendix 16. As a result, the implementation of enhanced digital assistance represents an independent intervention, in addition to the traditional PBL approach, in Experiment 5. Students who received the enhanced digital assistance intervention were designated as having $Digital$ equal to 1, while those who did not receive the intervention were designated as having $Digital$ equal to 0.

Experiment 6 centers on the examination of the contextual variable of pro-PBL family cultural environment, which is represented as a binary variable $ProPBL$. This variable reflects the prior overseas education experience of either the father or mother of the student. If either parent has had prior overseas education experience, the variable takes the value of 1, while a value of 0 is assigned in the absence of such experience. The term 'overseas' refers to regions beyond mainland China, as defined for the purpose of capturing the pro-family cultural environment discussed in Section 2.5.3. The coding of this binary contextual variable was based on the educational background of the student's parents, as provided by the RZS School.

4.3 Statistical analysis models

The statistical analysis models used in the thesis can be grouped into four categories: one-way variance analysis, multi-way variance analysis, linear regression model, and path analysis upon a model with mediating factor. Variance analysis, in turn, encompasses a broad range of techniques including Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA), Analysis of Covariance (ANCOVA), Multivariate Analysis of Covariance (MANCOVA), and multi-way ANOVA/ANCOVA. The programming language utilized in this thesis for the statistical analysis is R, with further details on the R version and packages used provided in Appendix 17.

4.3.1 One-way variance analysis models

In the realm of PBL-related experimental studies, variance analysis is a widely utilized statistical analysis model, with one-way ANOVA/MANOVA being one of the most frequently used models (Barth et al., 2019; O. Chen et al., 2016, 2019, 2020, 2021; Hsu et al., 2015; e.g., Klahr and Nigam, 2004; Kyun et al., 2013; Likourezos and Kalyuga, 2017; Matlen and Klahr, 2013; Moreno, 2004; Nachtigall et al., 2020; Steenhof et al., 2020; Zambrano et al., 2019b, 2019a). Although less frequently employed than ANOVA and MANOVA, ANCOVA and MANCOVA have been also used in some previous PBL-related experimental studies (Ajai et al., 2013; Argaw et al., 2016; Chase and Klahr, 2017; Firdaus and Herman, 2017; Loibl et al., 2020; e.g., Rittle-Johnson, 2006; L. Zhang, 2018, 2019). One-way variance analysis is mainly utilized in Experiment 1 of my thesis.

The philosophy of ANOVA in treatment experiment is to constitute a ratio of the explained variance over unexplained variance, taking into account their respective degrees of freedom. This ratio is expected to conform to an F-distribution with the given degrees of freedom, provided that the treatment grouping does not result in a reduction of variance. It can be mathematically expressed as:

$$F = \frac{DeV_T/(g-1)}{RSS_T/(n-g)} \qquad (4.2)$$

where DeV_T represents the variance between groups, or the portion of variance that is attributed to the grouping established in the experiment. On the other hand, RSS_T is defined as the variance within groups, or the portion of variance that is not explained by the experimental grouping. The variables g and n correspond to the number of experimental groups and the total number of participants involved in the study, respectively.

The numerator and denominator of equation (4.2) are both scaled χ^2 statistics. If the treatment grouping does not additionally reduce the variance, the F value in equation (4.2) should adhere to the $F(k-1, n-g)$ distribution. Otherwise, the numerator will be scaled by a greater[27] amount than the denominator. Hence, a one-tailed F test can be performed to test the null hypothesis that the treatment grouping contributes to the explanation of the variance in response variables.

Compared to ANOVA, MANOVA is employed when the response variables are not assumed to be independent from one another (O'Brien and Kaiser, 1985), which is reasonable in the context of the current thesis, particularly for the response variables $IScore$ and $DScore$. Although the procedure for calculating the outcome of the MANOVA is more intricate, the final outcome of the MANOVA is still an approximate F statistic. Further information regarding the calculation procedure can be found in Appendix 17.

ANCOVA or MANCOVA are extensions of ANOVA and MANOVA that allow for the examination of treatment effects while controlling for the influence of covariates (Keselman et

[27] It is because the variance explained by the treatment grouping cannot be negative.

al., 1998). These extensions can be understood as partitioning the explained and unexplained variance after excluding the portion that is explained by covariates in a linear model. Given that stratified random sampling has been performed, as outlined in Section 4.1, the utilization of ANCOVA serves as a supplementary technique. ANCOVA in an experimental design can be mathematically expressed as:

$$F = \frac{(DeV_T - DeV_{CT})/(g-1)}{(RSS_T - DeV_{CE})/(n-k-g)} \qquad (4.3)$$

where the numerator corresponds to the adjusted between-group variance DeV_T, taking into account the between-group variance DeV_{CT} that is attributed to covariate differences across groups. The denominator, on the other hand, reflects the adjusted within-group variance RSS_T, considering the within-group variance DeV_{CE} that is attributed to covariate differences across groups. The variables g and n denote the number of experimental groups and the total number of participants involved in the study, respectively. The variable k is the amount of covariates.

4.3.2 Multi-way variance analysis models

When the interest of research is to analyze the combined impact of multiple independent treatments (Experiments 2, 4, and the digital-assistant part of Experiment 5 in this thesis) or multiple independent variables (Experiments 3, 6, and the previous-PBL-experience part of experiment 5 in the thesis), multi-way variance analysis is employed to determine the significance of such interactive effects. Previous studies within the realm of PBL have also utilized multi-way variance analysis as demonstrated in works such as O. Chen et al. (2016), O. Chen et al. (2019), Ashman et al. (2020), O. Chen et al. (2020), and O. Chen et al. (2021).

Multi-way variance analysis assesses the deviation that results from the interaction of multiple classifications (Fujikoshi, 1993). For instance, when using two-way classification, the F statistic of the interaction term can be represented as:

$$F = \frac{(DeV_{T2} - DeV_{T1})/(g_2 - g_1)}{RSS_{T2}/(n-g_2)} \qquad (4.4)$$

where the numerator $DeV_{T2} - DeV_{T1}$ is the difference in variances between models with and without considering the interaction term. RSS_{E2} is residual sum of squares after two-way classification. $g1$ and $g2$ represent the number of groups under one-way and two-way

classification, respectively, and n is the total number of participants. The F value in equation (4.4) should conform to the $F(g2-g1, n-g2)$ distribution.

Three-way variance analysis, although mathematically more complex, follows the same basic principle as two-way variance analysis. Further information on this topic can be found in Kiers and Mechelen (2001) and Dawson and Richter (2006). In this thesis, three-way variance analysis was applied in Experiment 2 to study the synergetic effect of PBL elements, while two-way variance analysis was utilized in Experiments 3 to 6 to evaluate the influence of contextual factors on the effectiveness of PBL.

4.3.3 Linear regression model

Although most PISA and TIMSS studies adopt a linear model analysis approach, relatively few PBL experimental studies utilize linear model analysis. Nonetheless, there have been a great number of PBL experimental studies that employ the two-sample t-test analysis method (Aidoo et al., 2016; Amalia et al., 2017; Ashman et al., 2020; De Witte and Rogge, 2016; Hendarwati et al., 2021; Hendriana et al., 2018; Imanieh et al., 2014; Kazemi and Ghoraishi, 2012; Penjvini and Shahsawari, 2013; e.g., Stull and Mayer, 2007; Westhues et al., 2014), which is essentially equivalent to a linear model with a single binary independent variable.[28]

This thesis employs linear regression models with interaction terms in Experiments 2 to 6. When the estimation are based on the assumption that the residuals follow a Gaussian distribution, the stepwise linear regression model with interaction term is similar to multi-way ANCOVA (Keselman et al., 1998). The linear regression model provides an intuitive demonstration of the magnitude of difference between experimental groups, which cannot be achieved through variance analysis without conducting *post hoc* analysis.

Thus, the three-way classification of Experiment 2 can be analyzed by the following linear equation:

[28] While the t-statistic generated from the two-sample t-test analysis is equivalent to the square root of the F statistic obtained from ANOVA when the treatment variable is binary, the conceptual basis of the two-sample t-test analysis aligns more closely with the linear regression model. This is because the linear regression model estimates the coefficients while assuming the distribution of residuals, as opposed to the ANOVA-family analysis approach which compares variances.

$$
\begin{aligned}
Response_i \quad &= \beta_0 + \beta_1 ModGuid_i + \beta_2 LateGuid_i + \beta_3 Collective_i \\
&+ \beta_4 ModGuid_i * LateGuid_i + \beta_5 ModGuid_i * Collective_i \\
&+ \beta_6 LateGuid_i * Collective_i \\
&+ \beta_7 ModGuid_i * LateGuide_i * Collective_i \\
&+ \sum_{k=1}^{4} \gamma_k Z_{ki} + \epsilon_i \qquad (4.5)
\end{aligned}
$$

where i denotes each student. The response variable, $Response$, could represent $IScore$, $DScore$, $Cogload$, $Enjoy$, or $SelfEff$ as indicated in Equation 4.5. The covariate variables, Z_{ki}, include Age, $Gender$, RPF, and $SESI$ as specified in Equation 4.5. $ModGuid$, $LateGuid$, and $Collective$ represent PBL pedagogical elements, and their definitions and constructions can be found in Table 4.2.

The coefficients β_1, β_2, and β_3 are assigned to capture the individual effect of each of the PBL pedagogical elements. The coefficients β_4, β_5, and β_6 capture the combined effect of any two PBL elements. β_7 is assigned to capture the combined effect of all three PBL elements. Together, β_4, β_5, β_6, and β_7 encapsulate the synergistic impact of PBL, which is the focus of Experiment 2. No predictions are made regarding the signs of β_4, β_5, β_6, or β_7 in Section 3.1.2.

A general form of linear models used in Experiments 3 to 6 could be outlined as:

$$
\begin{aligned}
Response_i \quad &= \beta_0 + \beta_1 PBL_i + \beta_2 Contextual_i \\
&+ \beta_3 PBL_i * Contextual_i + \sum_{k=1}^{4} \gamma_k Z_{ki} + \epsilon_i \qquad (4.6)
\end{aligned}
$$

where i represents each student. The response variable, $Response$, can be one of $IScore$, $DScore$, and $Cogload$. The covariate variables, Z_{ki}, encompass Age, $Gender$, RPF, and $SESI$ as indicated in Equation 4.6. PBL stands for the overall PBL treatment, and $Contextual$ is the contextual variable which could be $Hgrade$, $Complex$, $CumuPBL$, $Digital$, or $ProPBL$ in different experiments. Consequently, the coefficients β_1 and β_2 of Equation 4.6 capture the effects of PBL and $Contextual$ respectively, while the coefficient β_3 of Equation 4.6 captures the interactive effect between PBL and $Contextual$. The variable of interest is β_3, which examines the effect of contextual factor on PBL efficacy.

The estimation of the linear regression models presented in Equations 4.5 and 4.6 may be subject to one potential issue. The commonly employed estimation technique, such as OLS estimation, assumes that the residual ϵ_i follows a Gaussian distribution. However, the dependent variables in Equations 4.5 and 4.6 are always bounded and non-negative, such as the $IScore$ variable which can only take values ranging from 0 to 100. This could result in a violation of the linear regression model's assumption regarding the error term distribution. To address this issue, a logit transformation can be applied to the dependent variable (Lesaffre et al., 2007). For instance, the $IScore$ variable can be transformed to:

$$ln\_IScore = ln(\frac{IScore}{100 - IScore}) \qquad (4.7)$$

where the new dependent variable, ln_IScore, is non-bounded and derived as the natural logarithm of the ratio between the original $IScore$ and $100 - IScore$. The linear regression analysis was conducted in Experiments 2 through 6 using logit-transformed response variables in addition to the raw variable.[29] The untabulated results of the logit-transformed regression affirm the robustness of the conclusion drawn in this thesis.

4.3.4 Path analysis

The linear regression in the previous section is only apt for testing a model with single endogenous outcome. However, as illustrated in Figures 3.1 and 3.2, cognitive load, serving as the mediating factor between PBL and learning outcome, is also an endogenous variable in the model. By means of linear regression analysis, the impact of PBL and other contextual factors on both learning outcome and cognitive load can only be analyzed separately. Path analysis, however, provides the means to examine these causality paths simultaneously, thereby enabling differentiation between PBL's direct impact on learning outcome and its indirect impact through cognitive load. Taking Figure 3.1 as an example, the path 'B' is the direct impact of PBL and the path 'A1->A2' is the indirect impact of PBL.

[29] The logit-transformation applied in Section 4.7 may not be suitable for the data under consideration, as it could result in a skewing of the transformed $IScore$ and $DScore$ towards the right. However, it has been utilized as a robust test.

Path analysis, as with CFA, is a component of SEM. Path analysis is founded on the path tracing principle established by Wright (1960) and Stage et al. (2004) describe the application of path analysis in the field of education research. Certain studies related to PBL, such as the work of Liou (2021), have employed path analysis to investigate the indirect effect of inquiry-based learning through students' enjoyment. This thesis focuses on the indirect effect of PBL on students' cognitive load in Experiments 3 to 6. The standard errors of path analysis in this thesis were estimated through within-sample bootstrapping. For more details, see Appendix 18.

4.4 Experimental procedures and sample descriptions

Having previously discussed the grouping strategy, measurement construct, and statistical analysis models, this section revisits Table 4.1. This section proceeds to document the detailed experimental process and provides a thorough description of the samples that were eventually formed.

4.4.1 Experimental procedures

The first experiment was conducted at LHZ School in the second semester of the academic year 2015-2016. The participants were selected from 8th grade students, numbering a total of 528. In accordance with Table 4.3, seven experimental groups, each consisting of 50 students, were established through the stratified random sampling method outlined in Section 4.1. The process of stratified random sampling was initiated by selecting students with the highest expected learning outcomes,[30] resulting in the exclusion of those with the lowest expected learning outcomes. The selected students received the pedagogical treatment described in Table 4.3 and took an immediate test on Newton's law of motion and answered questions related to cognitive load, enjoyment, and self-efficacy. Seven days later, the same students took a delayed test on Newton's law of motion. As no students missed the delayed test, the final sample of Experiment 1 is 350.

As depicted in Table 4.1, Experiment 1 aims to address the first two research questions and specifically to evaluate the effect of PBL pedagogical elements and the overall PBL approach as

[30] The results of the regression analysis for determining the expected learning outcome can be located in Appendix 5.

compared to traditional didactic teaching. To this end, a series of one-way variance analyses, as previously outlined in 4.3, were conducted to determine if the pedagogical treatment groups in Experiment 1 significantly reduce the variances in the response variables, which are $IScore$, $DScore$, and $CogLoad$ for RQ1, and $Enjoy$ and $SelfEff$ for RQ2. As these one-way variance analyses do not provide insight into which groups have better learning efficacy compared to the traditional didactic teaching group, a *post hoc* analysis, specifically Tukey's range test (Abdi and Williams, 2010), was also performed in Experiment 1.

The second experiment was conducted at LHZ School a year after Experiment 1. The participants, who were 8^{th} grade students, were selected from the 477 students enrolled in the school for the academic year 2016-2017. Experiment 2 aimed to examine the synergistic effect of PBL, which necessitated the use of a three-way variance analysis and a linear regression model with an interaction term, as outlined in Section 4.3. Thus, 400 participating students were required to be assigned to eight experimental groups: Didactic teaching, Moderate guidance, Late guidance, Collective learning, Moderate guidance and Late guidance, Moderate guidance and Collective learning, Late guidance and Collective learning, and Moderate guidance and Late guidance and Collective learning. The procedures for selecting and assigning students, as well as the testing process, were similar to those in Experiment 1. As there were no students who missed the delayed test, the final sample for Experiment 2 comprised 400 students.

The third experiment was conducted at the RZS School during the autumn semester of 2016. The participants for this experiment were drawn from both eighth and ninth grade students. RZS School, being a private institution, had smaller class sizes and fewer students compared to LHZ School. At the time of the experiment, the eighth grade consisted of 209 students and the ninth grade had 208 students. The objective of Experiment 3 was to investigate the contextual effect of $Hgrade$ and, thus, a two-way variance analysis and linear regression model with interaction terms were used, as specified in Section 4.3. For this experiment, 200 students were randomly assigned to four groups: non-PBL & Grade 8, PBL & Grade 8, non-PBL & Grade 9, and PBL & Grade 9. Unlike Experiment 1 and 2, questions about Enjoyment and Self-efficacy were not asked of the students as they were outside the scope of research question 3. There was one student who took the immediate test but did not complete the delayed test; thus, the final sample for Experiment 3 consisted of 199 students.

The fourth experiment was carried out at RZS School, taking place one year after Experiment 3. The objective of Experiment 4 was to examine the impact of the contextual factor of $Complex$, requiring the utilization of a two-way variance analysis and linear regression model with interaction terms, similar to the methodology employed in Experiment 3. As specified in Section 4.2, $Complex$ pertains to learning both Newton's laws and the law of conservation of energy. Consequently, 200 students were selected and randomly assigned to four experimental groups: non-PBL and non-complex task, PBL and non-complex task, non-PBL and complex task, and PBL and complex task. Similar to Experiment 3, questions regarding Enjoyment and Self-efficacy were not included. Two students took the immediate test but did not complete the delayed test, resulting in a final sample size of 198 students.

The fifth experiment moved back to LHZ School in December of 2018 and had two objectives. The first was to investigate the contextual effect of previous PBL experience ($CumuPBL$) and the second was to investigate the effect of digital assistance ($Digital$). So Experiment 5 contains two parts: Experiment 5A for the first research objective and Experiment 5B for the second. The two-way variance analysis and linear regression model with interaction terms were conducted separately for Experiments 5A and 5B. In the academic year of 2018-2019, LHZ School had ten 8th grade classes with a total of 453 students. During the autumn semester of 2018, two of these classes were chosen as the pilot for a learning innovation program that emphasized student-centered learning, as encouraged by the local education department of Guangxi province. These two classes were taught using principles that were closely aligned with PBL, providing their students with three months of previous PBL experience before they participated in Experiment 5A.

For Experiment 5A, two groups of 50 students were required, one for a non-PBL treatment group and one for a PBL treatment group, with some students coming from the three pilot classes and others from the remaining classes. As outlined in Section 4.1, the stratified random sampling was carried out separately for pilot and non-pilot classes to ensure that the experimental groups had a balance of students from both pilot and non-pilot classes. Experiment 5B required four experimental groups: non-PBL and non-digital treatment, PBL and non-digital treatment, non-PBL and digital treatment, and PBL and digital treatment. The definition and construction of digital treatment can be found in Section 4.2. Out of the total 453 8^{th} grade students, 300

students were selected for Experiment 5. As RQ4 did not pertain to long-term learning outcome, Enjoyment and Self-efficacy were not included in the experiment. No students missed the delayed test, so the final sample size was 300 students.

The sixth experiment was conducted at the RZS School, one year after the completion of Experiment 4. Its objective was to investigate the impact of the contextual factor $ProPBL$ on learning outcomes. The methodology employed for Experiment 6, including a two-way variance analysis and linear regression model with interaction terms, was similar to that used in Experiments 3 to 5. The construct of $ProPBL$, which pertains to a family cultural environment with a parent having prior overseas education experience, is defined in Section 4.2. The sample size for Experiment 6 was 100 grade-8 students, divided into two groups of 50 each, with the half students coming from Pro-PBL families and the others not. As described in Section 4.1, the stratified random sampling procedure was conducted separately for Pro-PBL families and non-Pro-PBL families to ensure that the experimental groups were composed of a balanced number of students from both Pro-PBL and non-Pro-PBL families. All students completed both the immediate and delayed tests, resulting in a final sample size of 100.

Table 4.5 provides a summary of the sample formulation across all experimental procedures discussed above.

Table 4.5: Sample formulation

Experiment	Time	# of groups	School	Grade	# of classes	# of total students	# of students attending experiment	# of students completing experiment
1	March 2016	7	LHZ School	Grade 8	11	528	350	350
2	February 2017	8	LHZ School	Grade 8	10	477	400	400
3	September 2016	4	RZS School	Grade 8	9	209	100	100
				Grade 9	9	208	100	99
4	September 2017	4	RZS School	Grade 8	10	223	200	198
5	December 2018	6	LHZ School	Grade 8	10	453	300	300

Experiment	Time	# of groups	School	Grade	# of classes	# of total students	# of students attending experiment	# of students completing experiment
6	September 2018	2	RZS School	Grade 8	10	236	100	100
				Total	69	2,334	1,550	1,547

Source: Compiled by author

4.4.2 Sample descriptions

In this section, I briefly describe the sample data formulated by the experimental procedures above. The means and standard deviations of the covariates for every experimental group of all the six experiments can be found in Table 4.6.

Table 4.6: Covariates across experimental groups

Experiment	Group	# of students	Male students (%)	Age (years)	SESI	RPF
1	*Collective*	50	54.0	13.96 (0.31)	5.99 (0.90)	72.18 (5.36)
	Didactic	50	54.0	13.96 (0.31)	5.84 (0.80)	72.82 (4.97)
	LateGuid	50	54.0	13.96 (0.31)	5.67 (0.81)	73.28 (4.86)
	MinGuid	50	54.0	13.96 (0.31)	5.92 (0.85)	73.86 (5.81)
	ModGuid	50	52.0	13.94 (0.31)	5.76 (0.77)	73.38 (5.32)
	PBL	50	52.0	13.94 (0.31)	5.81 (0.79)	72.36 (6.93)
	Placebo	50	54.0	13.96 (0.31)	5.88 (0.89)	73.12 (5.67)
2	*Collective*	50	44.0	13.84 (0.33)	6.35 (0.77)	75.06 (4.80)
	Didactic	50	54.0	13.88 (0.28)	6.23 (0.75)	74.70 (4.97)
	LateGuid	50	36.0	13.80 (0.31)	6.27 (0.75)	75.36 (4.73)
	LateGuid & Collective	50	42.0	13.82 (0.31)	6.22 (0.83)	75.72 (5.66)

Experiment	Group	# of students	Male students (%)	Age (years)	SESI	RPF
	ModGuid	50	40.0	13.90 (0.31)	6.36 (0.76)	74.14 (5.03)
	ModGuid & Collective	50	36.0	13.83 (0.31)	6.38 (0.68)	75.36 (5.29)
	ModGuid & LateGuid	50	56.0	13.91 (0.28)	6.21 (0.76)	75.14 (5.19)
	PBL	50	42.0	13.88 (0.31)	6.30 (0.84)	74.64 (4.70)
3	*Didactic*	50	50.0	13.51 (0.26)	8.14 (0.80)	80.70 (4.67)
	Didactic & Hgrade	50	60.0	14.54 (0.28)	8.04 (0.75)	80.08 (5.37)
	PBL	49	40.8	13.47 (0.31)	8.06 (0.75)	80.69 (4.17)
	PBL & Hgrade	50	48.0	14.47 (0.28)	7.92 (0.77)	79.52 (4.92)
4	*Didactic*	50	46.0	13.43 (0.32)	8.21 (0.77)	83.00 (4.11)
	Didactic & Complex	49	55.1	13.49 (0.32)	8.25 (0.80)	83.63 (3.93)
	PBL	50	44.0	13.38 (0.31)	8.06 (0.83)	82.72 (4.24)
	PBL & Complex	49	51.0	13.44 (0.30)	8.02 (0.62)	83.47 (3.84)
5A	*Didactic (CumuPBL=0)*	25	60.0	13.79 (0.30)	6.55 (0.69)	76.44 (5.09)
	Didactic (CumuPBL=1)	25	56.0	13.78 (0.36)	6.57 (0.79)	75.64 (5.15)
	PBL (CumuPBL=0)	25	40.0	13.73 (0.32)	6.58 (0.59)	77.20 (5.06)
	PBL (CumuPBL=1)	25	36.0	13.75 (0.30)	6.54 (0.94)	75.88 (4.10)
5B	*Didactic*	50	40.0	13.73 (0.30)	6.70 (0.86)	76.00 (4.42)
	Didactic & Digital	50	40.0	13.76 (0.30)	6.46 (0.66)	75.80 (3.94)
	PBL	50	50.0	13.78 (0.28)	6.52 (0.79)	74.72 (6.24)
	PBL & Digital	50	58.0	13.77 (0.29)	6.59 (0.77)	76.20 (5.40)

Experiment	Group	# of students	Male students (%)	Age (years)	SESI	RPF
6	*Didactic (ProPBL=0)*	25	48.0	13.50 (0.29)	8.20 (0.76)	85.44 (2.36)
	Didactic (ProPBL=1)	25	44.0	13.49 (0.33)	8.09 (0.83)	85.40 (3.06)
	PBL (ProPBL=0)	25	48.0	13.51 (0.30)	7.98 (0.95)	85.52 (2.99)
	PBL (ProPBL=1)	25	56.0	13.46 (0.33)	8.05 (0.63)	84.00 (4.41)

Source: Compiled by author

Table 4.6 displays information regarding several variables across seven experimental groups. The first column indicates the experiment number, while the second column provides the name of each experimental group. The third column lists the size of each experimental group, which is in accordance with Table 4.5. The fourth column provides the proportion of male students in each experimental group. The fifth column presents the mean and standard deviation of the students' age, with the mean indicated by the upper number and the standard deviation in parentheses. The mean and standard deviation of the $SESI$ and RPF variables are similarly presented in the sixth and seventh columns, respectively.

The data in Table 4.6 indicate that the gender distribution and average age of the students are relatively similar across the different groups within each experiment. On the other hand, the $SESI$ and RPF values exhibit significant differences between the groups. In particular, the students in Experiments 3, 4, and 6 showed significantly higher SESI values, which can be attributed to the fact that these experiments were conducted at RZS School, a private school located in a wealthy area of Zhejiang province, while the other experiments took place at LHZ School, a public school located in Guangxi province.

Since the stratified random sampling methodology described in Section 4.1 was employed, it is important to verify that there were no significant differences in the covariates among the experimental groups. In light of this, Table 4.7 presents the results of the ANOVA test for the covariates across experimental groups in all six experiments.

Table 4.7: Means and standard deviations of covariates across experimental groups

Experiment	RPF	SESI	Age	Gender
1	0.85 (0.36)	0.04 (0.83)	0.08 (0.78)	0.03 (0.87)
2	0.67 (0.41)	0.00 (0.99)	0.18 (0.67)	0.46 (0.50)
3	1.88 (0.17)	2.03 (0.16)	271.35*** (0.00)	0.17 (0.68)
4	0.82 (0.37)	0.66 (0.42)	0.42 (0.52)	0.67 (0.41)
5A	0.27 (0.61)	0.00 (0.99)	0.46 (0.50)	4.09** (0.05)
5B	0.27 (0.60)	0.58 (0.45)	0.30 (0.58)	1.94 (0.16)
6	1.00 (0.32)	0.73 (0.39)	0.05 (0.83)	0.35 (0.55)

Source: Compiled by author

Table 4.7 displays the results of the ANOVA test for the covariates across the experimental groups. The columns represent the covariates, while the rows indicate the specific experiments. Specifically, Experiment 5A and 5B are presented in separate rows. The numerical values presented in each cell represent the F-statistic of the ANOVA test, with the values in parentheses below indicating the corresponding p-value. The symbols *, **, and *** indicate significance levels of 0.1, 0.05, and 0.01, respectively.

The data presented in Table 4.7 indicates that, with the exception of the Age variable in Experiment 3 and the $Gender$ variable in Experiment 5A, there were no discernible differences in the covariates across the experimental groups. The former finding can be attributed to the higher age of higher-grade students in Experiment 3. It is important to note that the randomized assignment strategy appears to work as intended, as the RPF variable was found to be statistically non-significant in all experiments according to the ANOVA test results.

Following the completion of the six experiments, the response variable data were collected. Prior to presenting the results of the formal analysis in Chapter 5, a brief overview of the distribution of the response variables will be provided. Table 4.8 displays the descriptive statistics of the response variables.

Table 4.8: Descriptive statistics of response variables

Experiment	Variable	Mean	Median	Std.	P25	P75	N
1	IScore	67.29	65.00	7.63	65.00	70.00	350
	DScore	57.67	60.00	8.49	55.00	60.00	350
	CogLoad	0.18	0.31	0.89	-0.43	0.94	350
	Enjoy	-0.22	-0.26	0.82	-0.85	0.35	350
	SelfEff	-0.23	-0.43	0.84	-0.87	0.38	350
2	IScore	69.09	70.00	5.72	65.00	70.00	400
	DScore	59.21	60.00	6.53	55.00	60.00	400
	CogLoad	0.21	0.44	0.96	-0.43	1.05	400
	Enjoy	-0.05	-0.06	0.87	-0.67	0.72	400
	SelfEff	0.00	0.13	0.87	-0.67	0.75	400
3	IScore	75.28	75.00	8.38	70.00	80.00	199
	DScore	65.80	65.00	9.55	60.00	70.00	199
	CogLoad	-0.65	-1.17	0.88	-1.42	-0.07	199
4	IScore	68.74	70.00	9.26	65.00	75.00	198
	DScore	58.08	60.00	9.80	55.00	60.00	198
	CogLoad	0.53	1.05	0.99	-0.32	1.41	198
5A	IScore	73.40	70.00	8.79	68.75	80.00	100
	DScore	64.35	60.00	10.29	60.00	70.00	100
	CogLoad	-0.35	-0.43	0.87	-1.30	0.22	100
5B	IScore	73.50	70.00	9.39	65.00	80.00	200
	DScore	64.72	60.00	11.36	55.00	70.00	200
	CogLoad	-0.33	-0.43	0.82	-1.09	0.23	200
6	IScore	73.60	70.00	9.05	65.00	75.00	100
	DScore	63.15	60.00	9.15	60.00	65.00	100
	CogLoad	-0.25	-0.31	0.93	-1.21	0.31	100

Source: Compiled by author

Table 4.8 displays the mean, median, standard deviation, 25th percentile, and 75th percentile statistical values of the response variables. As previously discussed in Sections 3.1.2, 3.2.2, and 4.3, Experiments 1 and 2 include all five response variables, while Experiments 3 to 6 only include three response variables. The number of observations is consistent with Table 4.5.

The results presented in Table 4.8 indicate substantial variability in the dependent variables across the experiments. Specifically, the mean values of $IScore$ range from 67.29 to 75.28 and the standard deviation of these variables ranges from 5.72 to 11.36. It can be observed that the $IScore$ values are generally higher than the $DScore$ values, indicating that students tend to forget the knowledge they acquired over a seven-day period, thus performing worse in delayed tests. The $CogLoad$, $Enjoy$, and $SelfEff$ are CFA factors calculated based on corresponding questionnaire items (see Section 4.2), and their comparability across experiments is limited. Nonetheless, these factors exhibit relatively higher variability, as indicated by the ratio of standard deviation to mean, compared to $IScore$ and $DScore$.

4.5 Summary of Chapter 4

The study design, as presented in Chapter 4, consisted of six experiments aimed at addressing the research questions outlined in Chapter 3. The experiments were carried out at two different schools, over a period of two years, from March 2016 to December 2018, and involved students from grades 8 and 9. A total of 1,550 students were selected from a population of 2,334 students through the use of stratified random sampling based on predicted learning outcomes. The descriptive statistics, as presented in Table 4.7, indicate that the stratified random sampling used in this study effectively reduced disparity among the experimental groups.

To isolate the treatment effect, pre-recorded presentation videos were used and a placebo-controlled group was included. This effectively addressed the issue of potentially different learning environments among experimental groups, which has been a challenge faced by previous studies in this field. Cognitive load, students' enjoyment, and self-efficacy were measured through CFA of Likert-type questionnaires. The primary statistical models employed in this study include one-way variance analysis, multi-way variance analysis, linear regression, and path analysis. The empirical results of the research design and sample outlined in this chapter will be showcased in the next chapter.

5 Experimental findings

This chapter presents the main experimental findings of this thesis. The arrangement of the findings is predicated upon the order of the research questions that were posed. Section 5.1 outlines the empirical outcomes relative to RQ1, which encompass an evaluation of the short term efficacy of PBL components and the PBL approach as a whole, as well as an assessment of the synergistic effect of the PBL approach. Section 5.2 details the empirical outcomes related to RQ2, which focuses on the long-term indicators of PBL efficacy. Section 5.3 elucidates the empirical results pertaining to RQ3, with a focus on two contextual factors: the influence of students' prior knowledge and the effect of task complexity on learning. Section 5.4 presents the findings obtained from RQ4, which delves into three contextual factors: the impact of prior experience with PBL, the influence of the digital assistant environment, and the effect of a pro-PBL family culture. Finally, Section 5.5 summarizes this chapter by comparing the empirical findings to the predictions outlined in Chapter 3.

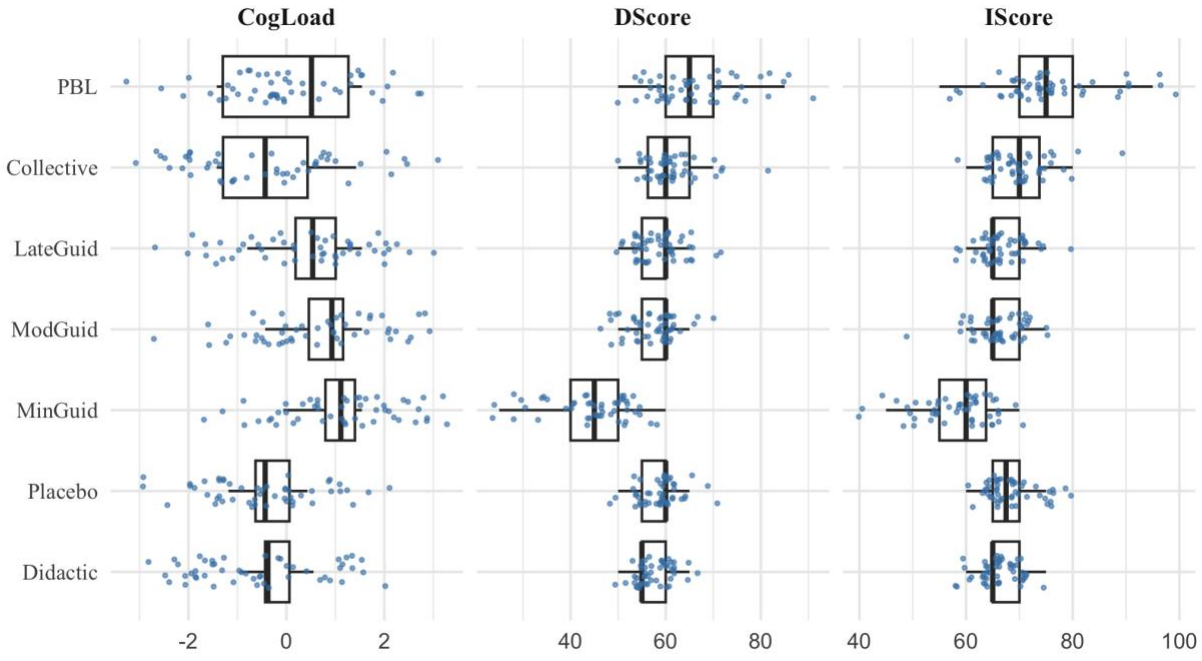
5.1 Empirical findings for RQ1

As previously discussed in Chapters 3 and 4, RQ1 is addressed through Experiments 1 and 2. Experiment 1 primarily evaluates the effectiveness of the elements of PBL and the holistic PBL approach, while Experiment 2 investigates the synergistic impact of the PBL approach.

5.1.1 Efficacy of PBL elements and overall PBL approach

Prior to conducting a formal one-way analysis of variance, Figure 5.1 presents a visual representation of the short-term learning outcomes across the seven experimental groups in Experiment 1. The vertical axis depicts the seven experimental groups, while the horizontal axis represents the three short-term response variables: $IScore$, $DScore$, and $CogLoad$. The figure displays the boxplots and jittered points for each experimental group, enabling a comparison of the response variables across the groups. The definitions and constructs of the response variables and treatment groups are described in Table 4.2.

Figure 5.1: Short-term learning outcomes across experimental groups



Source: Compiled by author

The results depicted in Figure 5.1 demonstrate the short-term learning outcomes across the various experimental groups. In particular, the analysis compares the seven groups in terms of student's testing scores and cognitive loads. It can be seen that there is a discernible trend in terms of cognitive load, where the Minimal guidance, Moderate guidance, Late guidance, and PBL groups show elevated levels of cognitive load relative to the didactic teaching group, aligning with the predictions of CLT. The Collective Learning and PBL group demonstrate large variances in cognitive load compared to other groups. The influence of the placebo effect on cognitive load was found to be insignificant.

In accordance with the prediction outlined in Section 3.1.2, Figure 5.1 shows that the PBL groups exhibited superior testing scores in comparison to the didactic teaching group. Conversely, the minimal guidance group students demonstrated inferior testing scores relative to the didactic teaching group, consistent with the prediction of an inverted U-shape relation between guidance level and learning outcomes. The placebo group also demonstrated a slightly higher $IScore$ relative to the didactic group, suggesting the presence of a potential positive

placebo effect. The comparison between the other groups and the didactic group is not clear based on the figure alone; therefore, it is necessary to perform more formal quantitative analysis.

5.1.1.1 One-way variance analysis

The first formal statistical analysis of RQ1 was performed using the datra from Experiment 1 through a series of one-way variance analyses as described in Section 4.3. Table 5.1 presents the results of the one-way variance analysis for RQ1.

Table 5.1: One-way variance analysis for RQ1

Variance analysis	IScore	DScore	CogLoad
ANOVA	42.01*** (0.00)	59.21*** (0.00)	27.44*** (0.00)
ANCOVA	42.93*** (0.00)	59.88*** (0.00)	26.97*** (0.00)
MANOVA	24.93*** (0.00)		
MANCOVA	24.76*** (0.00)		

Source: Compiled by author

Table 5.1 is organized with each row representing a type of one-way variance analysis. The columns represent the variables that are being tested. The MANOVA and MANCOVA analyses are performed on all three variables collectively, as opposed to a single variable. The values in the upper portion of each cell represent the F statistic or the approximate F statistic (for the MANOVA and MANCOVA analyses), while the values in parentheses denote the corresponding p-values. The symbols *, **, and *** indicate significance levels of 0.1, 0.05, and 0.01, respectively.

Table 5.1 reveals that, based on data collected from Experiment 1, the null hypothesis of the one-way variance analysis for RQ1 is rejected at a significance level of 1%. It indicates that the seven experimental groups in Experiment 1 significantly contribute to the variance of the three short-term learning outcome indicators. This conclusion is consistent with the observations presented in Figure 5.1. However, as previously discussed in Section 4.4, the one-way variance analysis does not allow for a comparison of the learning efficacy between the experimental groups and

the traditional didactic teaching group. To address this, a post-hoc analysis was conducted in Experiment 1, and its results are presented in the following section.

5.1.1.2 Post-hoc analysis

Table 5.2 displays the results of the post-hoc analysis for RQ1, obtained from Experiment 1. The results are derived from a Tukey's range test, a statistical method conceptually similar to a two-sample t-test. However, unlike the latter, the Tukey's range test conducts all pairwise comparisons simultaneously, as previously detailed in Section 4.4.

Table 5.2: Post-hoc test results of RQ1

Compare	IScore	DScore	CogLoad
Didactic-Collective	-3.80** (0.02)	-3.30* (0.09)	0.10 (0.99)
Didactic-LateGuid	-0.50 (1.00)	-1.20 (0.95)	-0.65*** (0.00)
Didactic-ModGuid	0.50 (1.00)	0.20 (1.00)	-1.02*** (0.00)
Didactic-Placebo	-2.30 (0.44)	-1.60 (0.84)	0.07 (1.00)
Didactic-PBL	-9.30*** (0.00)	-8.80*** (0.00)	-0.39 (0.13)
MinGuid-Didactic	-8.50*** (0.00)	-12.80*** (0.00)	1.25*** (0.00)

Source: Compiled by author

Table 5.2 presents the results of the pairwise comparisons between the didactic group and each of the other experimental groups. Each row in the table represents a comparison between two groups, with the difference in their mean values displayed in the upper portion of each cell. The values in parentheses represent the corresponding p-values. The symbols *, **, and *** indicate significance levels of 0.1, 0.05, and 0.01, respectively.

The results of the pairwise comparisons are consistent with the observations made from Figure 5.1. In particular, students in the PBL group showed significantly higher scores than those in the didactic group in both the immediate and delayed tests. This difference was statistically significant at the 1% level and was also significant in magnitude, with students in the PBL group

scoring 9.3 and 8.8 points higher in the immediate and delayed tests, respectively. Meanwhile, students in the Collective learning group also showed higher scores than those in the Didactic group, although to a lesser extent.

Conversely, students in the Minimal guidance group showed the lowest performance in both the immediate and delayed tests, suggesting that reducing guidance to a minimal level negatively impacts students' learning outcomes. The results support the notion that reducing guidance too far can have adverse effects on students' learning.

The results for cognitive load are also in line with the observations made from Figure 5.1. Students in the Minimal, Moderate, and Late guidance groups experienced higher cognitive load compared to those in the Didactic group. This aligns with the predictions of CLT, which posits that providing fewer or later guidance leads to higher cognitive load for students. Interestingly, students in the PBL and the Collective learning groups did not experience higher cognitive load compared to those in the didactic group. This contradicts the predictions of CLT that PBL would lead to higher cognitive load for students.

5.1.2 Synergistic effect of PBL approach

The supplementary inquiry of RQ1 explores the potential synergistic effect of the PBL approach on students' short-term learning outcomes. This examination is performed using the data collected in Experiment 2, as outlined in Chapter 4. The analysis of the data is conducted through a three-way variance analysis and a linear regression model with interaction terms.

5.1.2.1 Three-way variance analysis

Before conducting the regression analysis, the synergistic effect of PBL on students' short-term learning efficacy was initially explored through a three-way variance analyses, as outlined in Section 4.3. The three-way variance analysis also includes ANOVA, MANOVA, ANCOVA, and MANCOVA analyses, similar to those performed in the one-way variance analysis. The results of the three-way variance analysis for RQ1 are presented in Table 5.3.

Table 5.3: Three-way variance analysis for RQ1

Variance analysis	Interaction	IScore	DScore	CogLoad
ANOVA	Two-way	7.59*** (0.00)	4.92*** (0.00)	10.92*** (0.00)
	Three-way	3.60* (0.06)	3.33* (0.07)	0.47 (0.49)
ANCOVA	Two-way	7.36*** (0.00)	5.27*** (0.00)	10.26*** (0.00)
	Three-way	3.85* (0.05)	2.87* (0.09)	0.51 (0.48)
MANOVA	Two-way	4.98*** (0.00)		
	Three-way	1.85 (0.14)		
MANCOVA	Two-way	4.90*** (0.00)		
	Three-way	1.80 (0.15)		

Source: Compiled by author

Table 5.3 lists the specific method of variance analysis in the first column and the level of interaction in the second column. The remaining columns, 3 to 5, present the overall results for the three short-term learning outcome indicators. The values in the cells of the table indicate the F statistic or the approximate F statistic (for the MANOVA and MANCOVA analyses), while the corresponding p-values are provided in parentheses. The symbols *, **, and *** indicate significance levels of 0.1, 0.05, and 0.01, respectively.

The results of the analysis presented in Table 5.3 demonstrate that two-way interaction terms have a significant effect on the variances of all three short-term learning outcome indicators. This suggests that the combination of any two PBL pedagogical elements has a significant impact on students' short-term learning, although the direction of this impact is yet to be known. Additionally, the three-way interaction term also exhibits a marginally significant effect on $IScore$ and $DScore$ when analyzed through the ANOVA and ANCOVA models. These results may suggest the existence of a synergistic effect of the PBL approach, but further clarification through linear regression analysis is required to fully understand the influence at play.

5.1.2.2 Linear regression analysis

Table 5.4 presents the results of the linear regression analysis which aimed at exploring the synergistic effect of the PBL approach in terms of students' short-term learning efficacy. This analysis was performed using data collected from Experiment 2.

Table 5.4: Linear regression analysis for RQ1

Dependent Variable	**IScore**	**DScore**	**CogLoad**
Intercept	36.653*** (11.99)	54.402*** (13.84)	0.142 (2.06)
ModGuid	-0.393 (1.02)	-0.860 (1.18)	1.021*** (0.18)
LateGuid	-0.111 (1.02)	0.881 (1.18)	0.802*** (0.18)
Collective	2.077** (1.02)	3.126*** (1.18)	-0.124 (0.18)
ModGuid & LateGuid	-0.162 (1.45)	1.513 (1.67)	-0.698*** (0.25)
ModGuid & Collective	0.484 (1.44)	-0.493 (1.66)	-0.262 (0.25)
LateGuid & Collective	1.672 (1.44)	0.745 (1.66)	-0.222 (0.25)
ModGuid & LateGuid & Collective	4.000* (2.04)	3.987* (2.35)	-0.250 (0.35)
Adj. R squared	0.21	0.20	0.18
N	400	400	400

Source: Compiled by author

Table 5.4 presents the results of the linear regression analysis for PBL synergistic effect in RQ1. The independent variables in the linear model are listed in the first column (after the first row). Covariates including $Gender$, Age, $SESI$, and RPF are also included in the analysis but their estimates are omitted from Table 5.4 for conciseness. The results of the linear regression analysis with different dependent variables are displayed in columns 2 to 4. Specifically, column 2 presents the results for the model with $IScore$ as the dependent variable, column 3 for the model with $DScore$ as the dependent variable, and column 4 for the model with $CogLoad$ as the dependent variable. The estimated coefficients and their corresponding standard errors are

displayed in each cell, and their significance levels are indicated by asterisks, with *, **, and *** denoting significance levels of 0.1, 0.05, and 0.01, respectively.

Table 5.4 illustrates that the single element, $Collective$, has a positive impact on students' scores in both immediate and delayed exams. However, the single elements of $ModGuid$ and $LateGuid$ contribute to a higher cognitive load, but their influence on examination scores is insignificant. These findings align with the results of the one-way variance analysis from Experiment 1. The two-way interaction terms, as a whole, were shown to have a significant influence by the three-way variance analysis of Table 5.3. However, no single term of the two-way interaction was found to have a significant explanatory power for $IScore$ and $DScore$. The combination of moderate guidance and late guidance was found to significantly reduce cognitive load, while the three-way interaction term exhibited a marginally significant impact on short-term learning outcomes (i.e., at the 10% significance level). These findings provide evidence of a positive synergistic effect of PBL approach on short-term learning efficacy.

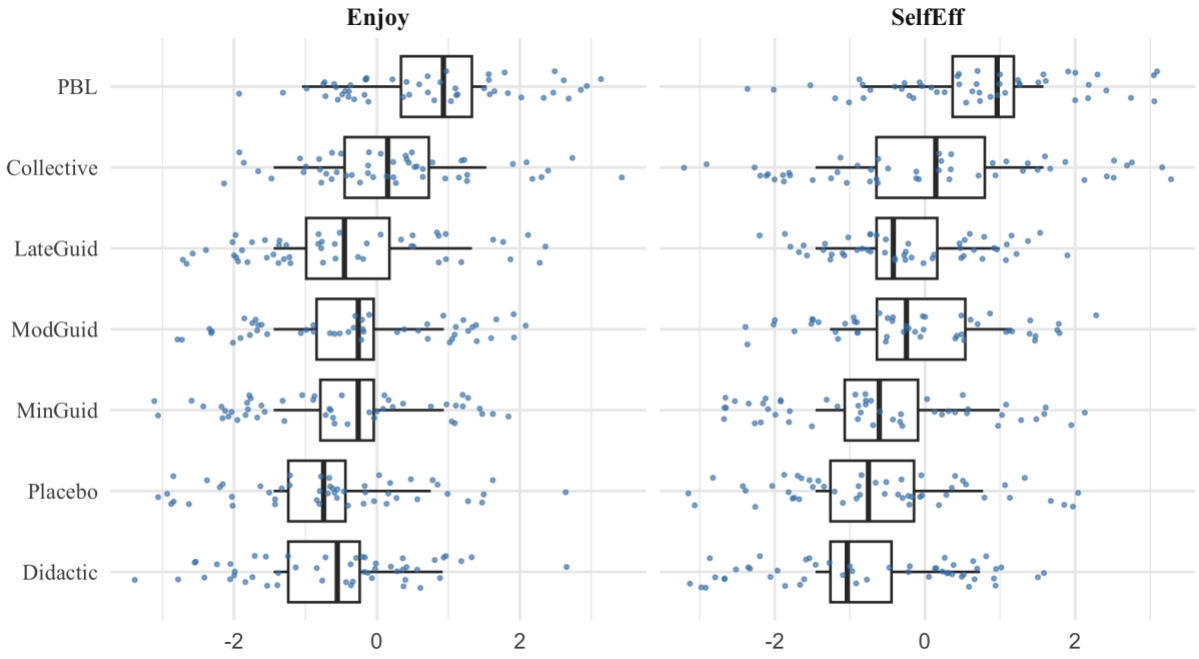
5.2 Empirical findings for RQ2

As was the case for RQ1, RQ2 was also addressed through the data collected from Experiments 1 and 2, with a focus on the two long-term learning indicators, $Enjoy$ and $SelfEff$. The findings of RQ2 are presented in two parts. The first part is concerned with the impact of the individual elements of PBL and the holistic PBL approach on the long-term learning indicators, which was primarily determined through Experiment 1. The second part focuses on the synergistic effects of the PBL approach on the long-term learning indicators, which was primarily evaluated through Experiment 2.

5.2.1 Efficacy of PBL elements and overall PBL approach

Figure 5.2 presents a visual representation of the long-term learning outcomes for the seven experimental groups in Experiment 1. The vertical axis displays the seven experimental groups, while the horizontal axis represents the two long-term response variables of interest, $Enjoy$ and $SelfEff$. The figure employs boxplots and jittered points to visually compare the response variables across the experimental groups, allowing for an initial examination of their distributions. The definitions and operationalizations of the response variables and experimental groups are detailed in Table 4.2.

Figure 5.2: Long-term learning outcomes across experimental groups



Source: Compiled by author

Figure 5.2 presents visual evidence indicating that the implementation of the PBL approach and its elements, particularly collective learning, have a positive impact on students' enjoyment and self-efficacy in Experiment 2. These results are consistent with the predictions of constructivism theory; however, a formal quantitative analysis in the form of a one-way analysis of variance is necessary to confirm these observations.

5.2.1.1 One-way variance analysis

The initial statistical evaluation of RQ2 was performed in Experiment 1 through the application of a series of one-way variance analyses, as outlined in Section 4.3. The one-way variance analysis suite, comprising ANOVA, MANOVA, ANCOVA, and MANCOVA analyses, was utilized. The results of the one-way variance analysis for RQ2 are presented in Table 5.5.

Table 5.5: One-way variance analysis for RQ2

Variance analysis	Enjoy	SelfEff
ANOVA	25.40*** (0.00)	27.11*** (0.00)

Variance analysis	Enjoy	SelfEff
ANCOVA	24.98*** (0.00)	27.22*** (0.00)
MANOVA	18.66*** (0.00)	
MANCOVA	18.52*** (0.00)	

Source: Compiled by author

Table 5.5 is structured with each row representing a different type of one-way variance analysis, including ANOVA and MANCOVA. The columns in the table denote the variables being tested, with the MANOVA and MANCOVA analyses performed collectively on both variables. The upper portion of each cell in the table displays the F statistic or the approximate F statistic (for the MANOVA and MANCOVA analyses), while the values in parentheses represent the corresponding p-values. Significance levels are indicated by the symbols *, **, and *** for 0.1, 0.05, and 0.01, respectively.

The results in Table 5.5 reveal that the null hypothesis of the one-way variance analysis for RQ2 was rejected at a significance level of 1% based on the data collected from Experiment 1. This suggests that the seven experimental groups in Experiment 1 significantly contribute to the variance of the two long-term learning outcome indicators, which is consistent with the observations shown in Figure 5.2. However, as previously discussed in Section 4.4, the one-way variance analysis does not provide a means to compare the long-term learning efficacy between the experimental groups and the traditional didactic teaching group. To address this limitation, a post-hoc analysis was conducted in Experiment 1 and its results are presented in a subsequent section.

5.2.1.2 Post-hoc analysis

Table 5.6 presents the outcomes of the post-hoc analysis for RQ2, obtained from the data collected in Experiment 1. The results are obtained through the application of Tukey's range test, the same method employed in Table 5.2.

Table 5.6: Post-hoc test results of RQ2

Compare	Enjoy	SelfEff
Didactic-Collective	-0.68*** (0.00)	-0.87*** (0.00)
Didactic-LateGuid	-0.19 (0.81)	-0.52*** (0.00)
Didactic-ModGuid	-0.27 (0.47)	-0.70*** (0.00)
Didactic-Placebo	0.07 (1.00)	-0.11 (0.99)
Didactic-PBL	-1.36*** (0.00)	-1.50*** (0.00)
MinGuid-Didactic	0.27 (0.46)	0.27 (0.47)

Source: Compiled by author

Table 5.6 presents the results of a pairwise comparison of the mean values between the Didactic group and each of the other experimental groups in Experiment 1. The results are displayed in terms of the differences in mean values and corresponding p-values. The significance levels are indicated by symbols *, **, and *** which correspond to 0.1, 0.05, and 0.01, respectively.

The results of the post-hoc analysis suggest that PBL and its elements, such as Collective Learning, can significantly enhance students' enjoyment and self-efficacy. In particular, the PBL and Collective Learning groups demonstrated a significantly higher level of $Enjoy$ and $SelfEff$ than the Didactic group, with a significance level of 1%. Additionally, both the Moderate Guidance and Late Guidance treatments were found to be effective in promoting students' self-efficacy. However, no significant effect was observed on students' enjoyment. In contrast, the Minimal Guidance treatment was not effective in enhancing either students' enjoyment or self-efficacy. This suggests that too little guidance may not only be detrimental to students' short-term learning efficacy, but also ineffective in promoting their long-term learning outcomes.

5.2.2 Synergistic effect of PBL approach

The supplementary investigation into RQ2 aims to assess the potential combined impact of the PBL approach on students' long-term learning outcomes. This examination was conducted using the data collected in Experiment 2, as described in Chapter 4. Similarly to RQ1, the analysis of

the data was performed through a three-way variance analysis and a linear regression model that includes interaction terms.

5.2.2.1 Three-way variance analysis

Prior to the execution of the regression analysis, the potential synergistic impact of PBL on students' long-term learning outcomes was preliminarily assessed through the use of three-way variance analyses, as described in Section 4.3. The findings of the three-way variance analysis regarding RQ2 are depicted in Table 5.7.

Table 5.7: Three-way variance analysis for RQ2

Variance analysis	Interaction	Enjoy	SelfEff
ANOVA	Two way	3.29** (0.02)	1.56 (0.20)
	Three way	2.94* (0.09)	1.33 (0.25)
ANCOVA	Two way	3.30** (0.02)	1.53 (0.21)
	Three way	2.83* (0.09)	1.71 (0.19)
MANOVA	Two way	2.42** (0.02)	
	Three way	2.17 (0.12)	
MANCOVA	Two way	2.41** (0.03)	
	Three way	2.31 (0.10)	

Source: Compiled by author

Table 5.7 provides evidence for the presence of a multi-way interactive effect of the PBL elements on students' enjoyment in learning. The results of the two-way interaction terms indicate a 5% level significant effect on the variance of $Enjoy$. The results of the three-way interaction terms, as determined through ANOVA and ANCOVA, indicate a marginally significant effect on the variance of $Enjoy$, with a significance level of 10%. However, the results of the three-way MANOVA and MANCOVA tests were not significant at the 1% level,

which may be influenced by including self-efficacy into the tests. These results suggest the presence of a synergistic effect of the PBL approach on students' enjoyment, but further clarification is necessary through linear regression analysis to fully comprehend the underlying influence.

5.2.2.2 Linear regression analysis

The results of the linear regression analysis aimed at investigating the synergistic impact of the PBL approach on students' long-term learning efficacy are presented in Table 5.8.

Table 5.8: Linear regression analysis for RQ2

Dependent Variable	**Enjoy**	**SelfEff**
Intercept	1.026 (1.78)	-2.846 (1.87)
ModGuid	0.217 (0.15)	0.343** (0.16)
LateGuid	0.358** (0.15)	0.167 (0.16)
Collective	0.602*** (0.15)	0.430*** (0.16)
ModGuid & LateGuid	0.124 (0.21)	-0.053 (0.23)
ModGuid & Collective	0.032 (0.21)	-0.141 (0.22)
LateGuid & Collective	-0.273 (0.21)	0.088 (0.22)
ModGuid & LateGuid & Collective	0.508* (0.30)	0.417 (0.32)
Adj. R squared	0.25	0.16
N	400	400

Source: Compiled by author

Table 5.8 illustrates the outcome of the linear regression analysis performed to investigate the synergistic effect of the PBL approach on students' long-term learning efficacy in accordance with RQ2. The independent variables in the model are listed in the first column, with covariates such as $Gender$, Age, $SESI$, and RPF also included but omitted from the table for brevity. The

results for the linear regression models with $Enjoy$ and $SelfEff$ as dependent variables are presented in columns 2 and 3, respectively. The estimated coefficients and their standard errors are displayed in each cell, and their level of significance is indicated through the use of asterisks, with *, **, and *** denoting significance levels of 0.1, 0.05, and 0.01, respectively.

Table 5.8 suggests that, when controlling for the interaction, the Collective learning component still has a significantly positive impact on both students' enjoyment and self-efficacy in learning. Moreover, the results indicate that Late guidance and Moderate guidance have positive impacts on students' learning enjoyment and self-efficacy respectively at a significance level of 5%. Although the multi-way variance analysis in Table 5.7 demonstrates the explanatory power of experimental grouping on students' learning enjoyment, no single two-way interaction term demonstrates significant results in Table 5.8, which may be for similar reasons as those indicated in Table 5.4. The three-way interaction term is marginally significant when the dependent variable is $Enjoy$, with a positive estimated coefficient, suggesting that the combination of all three PBL elements may enhance students' enjoyment of learning.

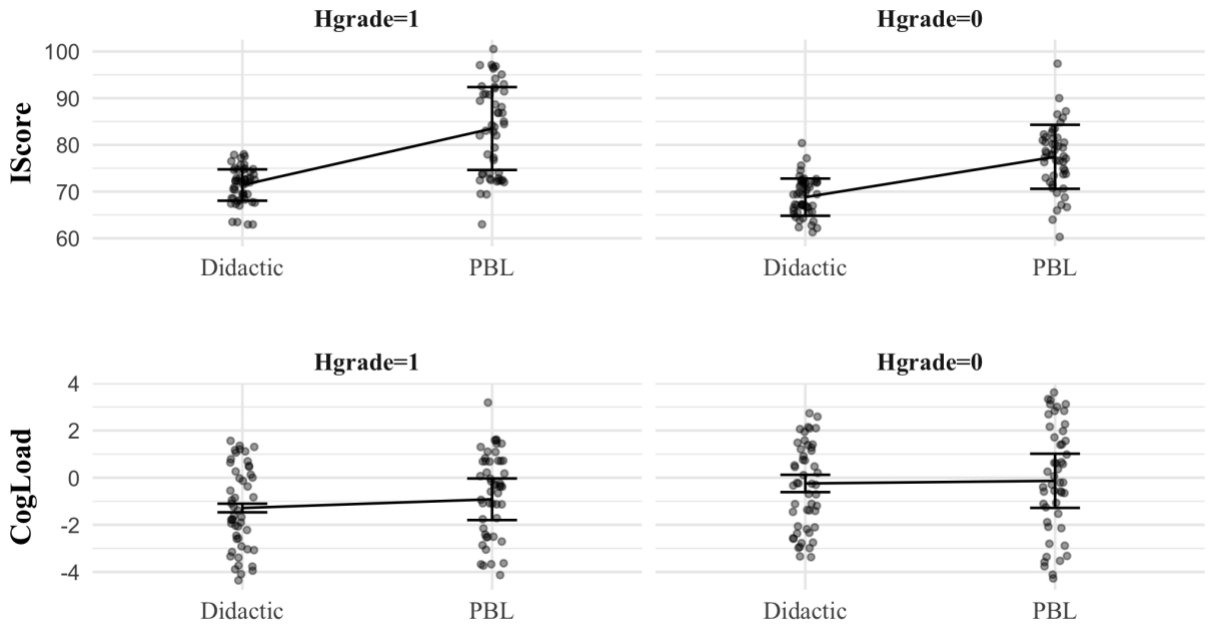
5.3 Empirical findings for RQ3

As previously mentioned in Chapter 3, RQ3 examines the impact of students' prior knowledge and the complexity of the learning task through Experiments 3 and 4. This section will first present the results regarding the contextual effect of students' prior knowledge.

5.3.1 Contextual effect of students' prior knowledge

Prior to conducting a formal quantitative analysis, Figure 5.3 provides a visual representation of the impact of students' prior knowledge on the PBL efficacy. The PBL treatment effects were depicted by connecting the error bars of the response variables for the PBL and didactic groups. For instance, taking the top-left portion of Figure 5.3 as an example, the error bars and jittered points above the 'Didactic' and 'PBL' labels represent the $IScore$ of Grade-9 students without and with the PBL treatment, respectively. The connecting line has a positive slope, indicating that when $HGrade$=1, the PBL treatment increases the students' immediate test scores. The top-right portion of Figure 5.3 follows the same approach, but for Grade-8 students. For the purpose of brevity, only the results of $IScore$ and $CogLoad$ are depicted in Figure 5.3.

Figure 5.3: PBL treatment effects by students' grades



Source: Compiled by author

Figure 5.3 depicts the results of Experiment 3, which aimed at exploring the contextual effects of students' prior knowledge on their learning outcomes. The figure presents a visual representation of the impact of PBL on the students' immediate test scores ($IScore$) and cognitive load ($CogLoad$) based on their grade. The results indicate that PBL leads to higher test scores for both grade 8 and grade 9 students. The positive impact of PBL on immediate performance appears to be more pronounced for higher-grade students. This pattern aligns with the prediction that expert learners can benefit more from PBL or are less likely to be negatively impacted by it. The slopes of PBL on cognitive load, as shown in the figure, were not found to be significantly different. Additionally, higher-grade students appear to have a lower level of cognitive load compared to lower-grade students, suggesting that learning the same physics knowledge is more challenging for lower-grade students.

5.3.1.1 Two-way variance analysis

As outlined in Section 4.3, the influence of the contextual factors on the efficacy of PBL can be examined through a two-way variance analysis. The results of this analysis with regards to the contextual effect of students' prior knowledge are presented in Table 5.9.

Table 5.9: Two-way variance analysis for contextual effects of students' grades

Variance analysis	Interaction	IScore	DScore	CogLoad
ANOVA	Two-way	3.88* (0.05)	1.25 (0.26)	1.47 (0.23)
ANCOVA	Two-way	4.18** (0.04)	1.31 (0.25)	1.36 (0.25)
MANOVA	Two-way	3.46** (0.02)		
MANCOVA	Two-way	3.51** (0.02)		

Source: Compiled by author

The results of the two-way variance analysis are presented in Table 5.9. The table displays the extent to which the combination of the PBL treatment and the contextual factor ($Hgrade$) contribute to the variance in the response variable. The cells of the table provide the F statistic or the approximate F statistic (for the MANOVA and MANCOVA analyses) with the corresponding p-values presented in parentheses. Significance levels are indicated by asterisks, with *, **, and *** representing 0.1, 0.05, and 0.01 respectively.

The results of the analysis, as shown in Table 5.9, reveal that the two-way interaction term between PBL and $Hgrade$ has a significant effect on the variance of $IScore$ with a significance level of 10% for ANOVA and 5% for ANCOVA. The interaction term, however, does not significantly reduce the variance of $DScore$, which is in line with the findings of previous studies reviewed in Section 2.3 that document the generally weakening efficacy of PBL. Additionally, the two-way variance analysis did not reject the null hypothesis for $CogLoad$, which contradicts the predictions of CLT. Further insights into this phenomenon will be explored through linear regression and path analysis in the following sections.

5.3.1.2 Linear regression analysis

The results of a linear regression analysis that explores the interaction effect between PBL and students' prior knowledge on learning efficacy are presented in Table 5.10. This analysis was conducted utilizing data collected from Experiment 3.

Table 5.10: Linear regression analysis for contextual effects of students' grades

Dependent Variable	IScore	DScore	CogLoad
Intercept	38.138* (21.45)	48.420* (25.76)	2.358 (2.60)
PBL	8.815*** (1.25)	9.981*** (1.50)	0.094 (0.15)
Hgrade	1.254 (2.06)	3.650 (2.48)	-0.924*** (0.25)
PBL & Hgrade	3.587** (1.75)	2.409 (2.11)	0.248 (0.21)
Adj. R squared	0.46	0.40	0.28
N	199	199	199

Source: Compiled by author

The independent variables in the model include PBL, $Hgrade$, and their interaction term. Covariates including $Gender$, Age, $SESI$, and RPF were also included in the analysis but are not reported in the table for conciseness. The estimated coefficients and their corresponding standard errors are displayed in each cell, and their significance levels are indicated by asterisks, with *, **, and *** denoting significance levels of 0.1, 0.05, and 0.01, respectively. The definitions and constructs of the response variables and treatment groups are given in Table 4.2.

The results of the linear regression analysis reveal that in Experiment 3 PBL is significantly and positively associated with both $IScore$ and $DScore$, supporting the findings of Experiment 1 reported in Section 5.1. The results indicate that PBL has no significant impact on students' cognitive load, which is consistent with the previous findings in Table 5.2 but contrary to the prediction of CLT. The analysis also shows that $Hgrade$ is negatively and significantly associated with cognitive load, suggesting that higher-grade students experience less cognitive challenge when learning new physics knowledge. However, their scores are not significantly higher even at a significance level of 10%.

The interaction term between PBL and $Hgrade$ is positively associated with $IScore$, indicating that for higher-grade students, PBL treatment can be more effective in promoting higher scores in the immediate test. However, the interaction term is not significantly related to $CogLoad$, suggesting that the additional benefits of PBL for higher-grade students are not due to reduced cognitive load. This will be further explored in the subsequent path analysis.

5.3.1.3 Path analysis

Table 5.11 presents the results of the path analysis conducted to examine the contextual effects of students' grades on their learning outcomes. As described in Section 4.3, path analysis enables us to examine complex models with multiple endogenous outcomes and distinguish between the direct and indirect effects of the independent variables of interest.

Table 5.11: Path analysis for contextual effects of students' grades

Path Analysis	Hgrade	PBL	Hgrade & PBL
Direct	-1.28 (1.97)	9.07*** (1.15)	4.27*** (1.63)
Indirect	2.53*** (0.84)	-0.26 (0.41)	-0.68 (0.59)
Total	1.25 (2.02)	8.82*** (1.22)	3.59** (1.72)

Source: Compiled by author

The columns in Table 5.11 represent the three variables of interest in Experiment 3, which are PBL, $Hgrade$, and their interaction term. The rows in the table indicate the direct, indirect, and total effects of these three variables on $IScore$. Although the results for other variables are not reported, it is important to note that the path analysis model controlled for those four covariates: $Gender$, Age, $SESI$, and RPF.

The direct effect refers to the extent to which the corresponding variable directly impacts $IScore$. The indirect effect refers to the extent to which the corresponding variable affects $IScore$ through its influence on $CogLoad$. The total effect is the sum of both the direct and indirect effects. The estimated coefficients and their corresponding standard errors are displayed in each cell, with the standard errors being calculated using a bootstrap technique (details of which can be found in Appendix 18). Significance levels of 0.1, 0.05, and 0.01 are indicated by asterisks – *, **, and *** respectively.

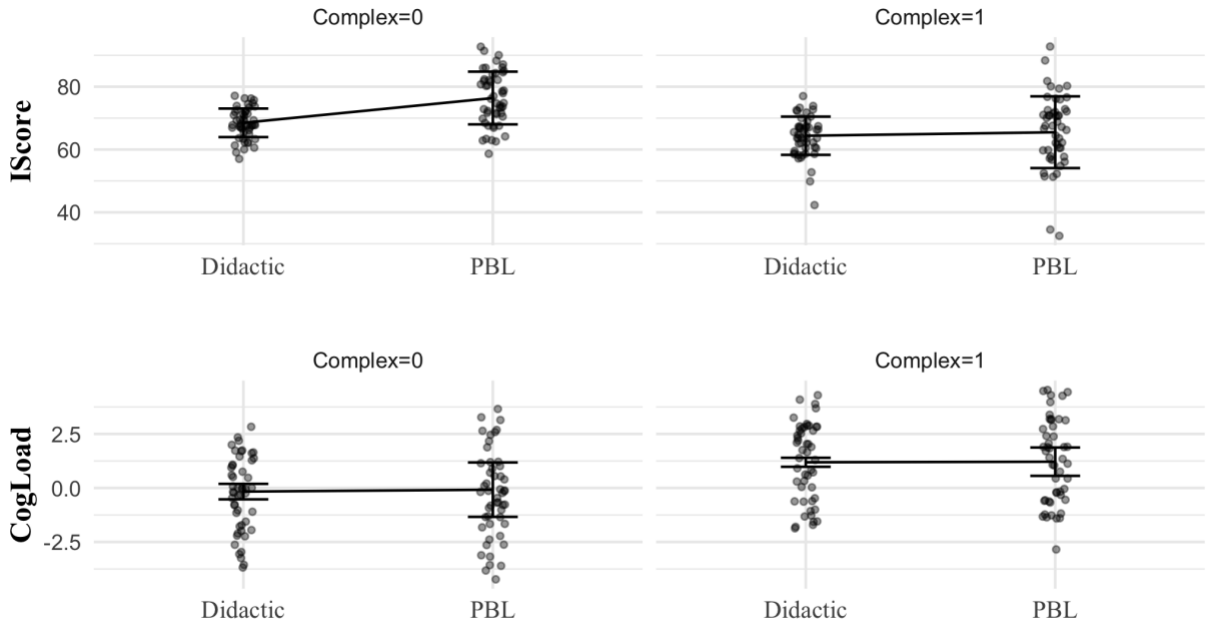
The results of the path analysis suggest that the positive impact of students' grades, represented by $Hgrade$, on their immediate learning efficacy, represented by $IScore$, is indirect and occurs through reducing their cognitive load, represented by $CogLoad$. Conversely, the positive impact of the PBL approach treatment, represented by PBL, is solely attributed to its direct effect on

$IScore$. Similarly, the total impact of the interaction term between PBL and $Hgrade$ is primarily driven by its direct effect on $IScore$. These results strengthen the inferences from the linear regression analysis, that the contextual effect of higher-grade students in Experiment 3 is mainly driven by the benefits beyond the channel through cognitive load.

5.3.2 Contextual effect of learning task complexity

The second research objective of RQ3 is to evaluate the contextual impact of task complexity on learning outcomes. Similar to the illustration presented in Figure 5.3, Figure 5.4 offers a visual representation of the relationship between task complexity and the effectiveness of the PBL approach. The connecting lines between the error bars indicate the PBL treatment effects for various levels of task complexity in Experiment 4.

Figure 5.4: PBL treatment effects by learning task complexity



Source: Compiled by author

Figure 5.4 reveals that the PBL treatment was effective in enhancing students' immediate test scores only when the learning task was of low complexity, characterized by fewer interactive elements. This is in line with the predictions of CLT. However, there is not a significant positive slope in the PBL treatment effect on $CogLoad$ when the complexity of the learning task

($Complex$) was equal to 1. This pattern deviates from the logic of CLT. Figure 5.4 also shows that the level of $CogLoad$ was higher for students facing more complex learning tasks, which are more likely to exhaust their working memory.

5.3.2.1 Two-way variance analysis

Following the visual comparison, Table 5.12 presents the results of the two-way variance analysis, which quantifies the contextual effect of learning task complexity.

Table 5.12: Two-way variance analysis for contextual effects of learning task complexity

Variance analysis	Interaction	IScore	DScore	CogLoad
ANOVA	Two-way	8.82*** (0.00)	8.55*** (0.00)	0.09 (0.77)
ANCOVA	Two-way	8.79*** (0.00)	8.65*** (0.00)	0.08 (0.78)
MANOVA	Two-way	6.68*** (0.00)		
MANCOVA	Two-way	6.72*** (0.00)		

Source: Compiled by author

The cells in Table 5.12 display the extent to which the combination of the PBL treatment and the contextual factor ($Complex$) contribute to the variance in the response variables. The F statistics or the approximate F statistics (for the MANOVA and MANCOVA analyses), along with the corresponding p-values, are provided in parentheses. Significance levels are indicated by asterisks, *, **, and ***, which denote 0.1, 0.05, and 0.01 levels of significance, respectively.

The results revealed that the interaction term between PBL and $Complex$ has a significant impact on the variance of $IScore$ and $DScore$, with a significance level of 1% as indicated by both ANOVA and ANCOVA. However, the interaction term was not found to significantly reduce the variance of $CogLoad$, which contradicts the predictions of CLT. Further investigation into this result will be explored through linear regression and path analysis in subsequent sections.

5.3.2.2 Linear regression analysis

The results of the linear regression analysis that investigates the interaction effect between the PBL treatment and learning task complexity on learning efficacy are presented in Table 5.13. This analysis is based on the data collected from Experiment 4.

Table 5.13: Linear regression analysis for contextual effects of learning task complexity

Dependent Variable	**IScore**	**DScore**	**CogLoad**
Intercept	13.294 (26.89)	-0.404 (27.04)	4.421* (2.49)
PBL	8.176*** (1.61)	8.467*** (1.62)	0.072 (0.15)
Complex	-4.505*** (1.62)	-6.279*** (1.63)	1.381*** (0.15)
PBL & Complex	-6.747*** (2.28)	-6.732*** (2.29)	-0.059 (0.21)
Adj. R squared	0.25	0.33	0.44
N	198	198	198

Source: Compiled by author

The model behind Table 5.13 includes the independent variables of interest, PBL, $Complex$, and their interaction term. The estimated coefficients, their corresponding standard errors, and their significance levels, indicated by asterisks, where *, **, and *** denote significance levels of 0.1, 0.05, and 0.01 respectively, are displayed in each cell. The definitions and constructs of the response variables and treatment groups can be found in Table 4.2.

The results of the linear regression analysis reveal that in Experiment 4, PBL is significantly and positively associated with both $IScore$ and $DScore$, confirming the findings from Experiment 1 reported in Section 5.1. The results also indicate that PBL has no significant impact on students' cognitive load, consistent with the previous findings in Table 5.2, but contrary to the predictions of CLT. Additionally, the results show that $Complex$ is positively and significantly associated with cognitive load, implying that learning tasks with higher levels of complexity are more likely to deplete students' working memory capacity. The significantly negative association between $Complex$ and test score further supports this notion.

The interaction term between PBL and $Complex$ is negatively associated with $IScore$ and $DScore$, indicating that the benefits of PBL treatment are offset to a great extent for learning tasks with higher complexity levels. However, the interaction term is not significantly related to $CogLoad$, suggesting that the diminished benefits of PBL for more complex learning tasks may not be due to increased cognitive load. This phenomenon will be further examined in subsequent path analysis.

5.3.2.3 Path analysis

The results of a path analysis aimed at exploring the contextual effects of learning task complexity on students' immediate learning efficacy (represented by $IScore$) are presented in Table 5.14. The methodology of path analysis, which is described in Section 4.3, allows one to assess complex models that incorporate multiple endogenous variables and to differentiate between the direct and indirect impacts of the independent variables of interest.

Table 5.14: Path analysis for contextual effects of learning task complexity

Path Analysis	Complex	PBL	Complex & PBL
Direct	2.57 (1.68)	8.54*** (1.39)	-7.05*** (1.96)
Indirect	-7.07*** (1.20)	-0.37 (0.75)	0.30 (1.06)
Total	-4.51*** (1.59)	8.18*** (1.58)	-6.75*** (2.23)

Source: Compiled by author

The analysis takes into consideration the impact of three variables of interest in Experiment 4 – PBL, $Complex$, and their interaction term – while controlling for the influence of four covariates: $Gender$, Age, $SESI$, and RPF.

The direct effect of a variable refers to its direct impact on $IScore$, while the indirect effect refers to its impact on $IScore$ through its influence on cognitive load ($CogLoad$). The total effect of a variable is the sum of its direct and indirect effects. The results are displayed in the table as coefficients with their corresponding standard errors. Significance levels of 0.1, 0.05, and 0.01 are indicated by asterisks *, **, and *** respectively.

The results of the path analysis indicate that the negative impact of learning task complexity ($Complex$) on students' immediate learning efficacy ($IScore$) is mainly driven by its indirect effect through an increase in cognitive load ($CogLoad$). On the other hand, the positive impact of the PBL approach treatment (PBL) is solely attributed to its direct effect on $IScore$. Similarly, the total negative impact of the interaction term between PBL and $Complex$ is mainly driven by its direct effect on $IScore$. These results further reinforce the inferences drawn from the linear regression analysis that the contextual effect of learning task complexity in Experiment 4 is primarily driven by costs beyond cognitive load.

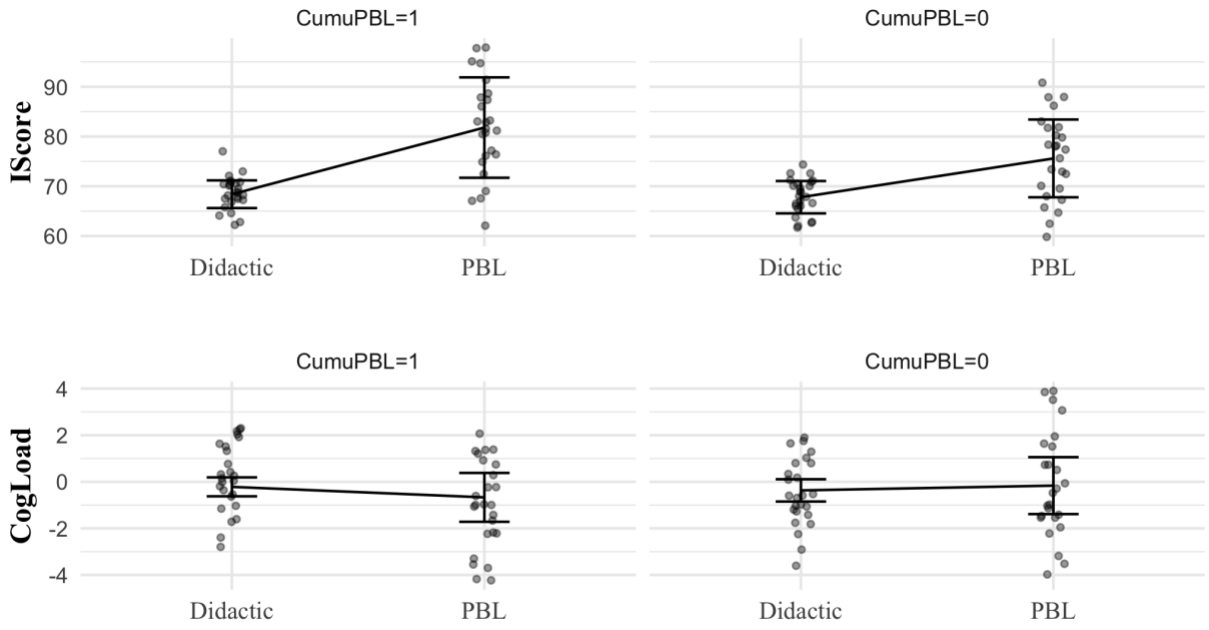
5.4 Empirical findings for RQ4

As previously discussed in Chapter 3, RQ4 evaluates the influence of students' prior PBL experience, utilization of digital guidance, and pro-PBL family culture via undertaking Experiments 5 and 6. This section will present the findings with respect to the contextual impact of students' previous PBL experience as the initial aspect for consideration.

5.4.1 Contextual effect of previous PBL experience

As previously demonstrated in Figures 5.3 and 5.4, Figure 5.5 provides a graphical representation of the connection between students' prior PBL experience and the efficacy of the PBL approach. The lines connecting the error bars illustrate the effects of the PBL treatment for students who have or have not previously engaged in a PBL experience in Experiment 5A.

Figure 5.5: Students' previous PBL experience and PBL treatment effects



Source: Compiled by author

Figure 5.5 provides a visual representation of the impact of PBL on students' immediate test scores ($IScore$) and cognitive load ($CogLoad$) based on the students' previous PBL experience, which was established through participation in a three-month pilot learning program prior to Experiment 5.

The results indicate that the PBL treatment was successful in enhancing students' immediate test scores, regardless of whether or not they had previous PBL experience. Additionally, the line connecting the Didactic and PBL groups for students with previous PBL experience appears to be steeper, suggesting that the positive impact of the PBL treatment may be more pronounced for these students.

In terms of students' cognitive load, although not prominently displayed in Figure 5.5, the results suggest that PBL treatment had a slight increase on cognitive load for students without previous PBL experience, while it had a slight decrease on cognitive load for students with previous PBL experience. This intriguing pattern warrants further quantitative analysis in the rest of this section.

5.4.1.1 Two-way variance analysis

The quantitative analysis of the contextual effect of students' previous PBL experience is presented in Table 5.15, which provides a two-way variance analysis.

Table 5.15: Two-way variance analysis for contextual effects of previous PBL experience

Variance analysis	Interaction	IScore	DScore	CogLoad
ANOVA	Two-way	4.33** (0.04)	5.33** (0.02)	3.60* (0.06)
ANCOVA	Two-way	4.20** (0.04)	5.13** (0.03)	3.27* (0.07)
MANOVA	Two-way	2.28* (0.08)		
MANCOVA	Two-way	2.20* (0.09)		

Source: Compiled by author

Table 5.15 shows the degree to which the combination of the PBL treatment and the contextual factor ($CumuPBL$) influence the variance of the response variables. The F statistics or the approximate F statistics (in the case of the MANOVA and MANCOVA analyses), along with their respective p-values, are presented in parentheses. Significance levels are indicated by *, **, and ***, which denote 0.1, 0.05, and 0.01 levels of significance, respectively.

The analysis results indicate that the interaction between PBL and $CumuPBL$ has a significant impact on the variance of $IScore$ and $DScore$ at the 5% level of significance as determined by both the ANOVA and ANCOVA analyses. Additionally, the interaction term was also found to significantly reduce the variance of $CogLoad$ at the 10% level of significance. This pattern was not observed in the contextual effects of $Hgrade$ and $Complex$ in this thesis. Further investigation into this result will be undertaken via linear regression and path analysis in subsequent sections.

5.4.1.2 Linear regression analysis

The findings from the linear regression analysis exploring the interaction effect of the PBL treatment and students' previous PBL experience on learning efficacy are presented in Table 5.16. This analysis was conducted using data obtained from Experiment 5A.

Table 5.16: Linear regression analysis for contextual effects of previous PBL experience

Dependent Variable	IScore	DScore	CogLoad
Intercept	12.070 (30.70)	29.693 (36.64)	6.095 (3.87)
PBL	7.858*** (1.93)	7.646*** (2.31)	0.180 (0.24)
CumuPBL	0.687 (1.91)	0.921 (2.28)	0.172 (0.24)
PBL & CumuPBL	5.532** (2.70)	7.295** (3.22)	-0.615* (0.34)
Adj. R squared	0.41	0.39	0.05
N	100	100	100

Source: Compiled by author

The model used in Table 5.16 includes the independent variables of PBL, $CumuPBL$, and their interaction term. The estimated coefficients, corresponding standard errors, and significance levels, denoted by asterisks, with *, **, and *** indicating significance levels of 0.1, 0.05, and 0.01 respectively, are displayed in each cell. The definitions and constructs of the response variables and treatment groups can be found in Table 4.2.

The results from the linear regression analysis reveal that in Experiment 5A, PBL was significantly and positively associated with both $IScore$ and $DScore$, consistent with the findings from Experiment 1 reported in Section 5.1. The results also indicate that PBL had no significant impact on students' cognitive load, which is consistent with the previous findings in Table 5.2 but contradicts the predictions of CLT. Additionally, the results show that $CumuPBL$ was not related to either test scores or cognitive load, indicating that students' previous PBL experience does not have a direct effect on the learning outcome in Experiment 5A.

The interaction term between PBL and $CumuPBL$ was positively associated with $IScore$ and $DScore$ at a 5% significance level, which suggests that the benefits of PBL treatment are more pronounced for students who had prior PBL experience through participation in the pilot program classes in Experiment 5A. Furthermore, the interaction term was negatively associated with $CogLoad$ at a 10% significance level, implying that the strengthened benefits of PBL for students with previous PBL experience may be partially due to reduced cognitive load.

5.4.1.3 Path analysis

Table 5.17 presents the findings of a path analysis designed to examine the impact of prior experience with PBL on students' learning outcomes.

Table 5.17: Path analysis for previous PBL experience

Path Analysis	CumuPBL	PBL	CumuPBL & PBL
Direct	1.37 (1.59)	8.57*** (1.61)	3.09 (2.28)
Indirect	-0.68 (0.92)	-0.71 (0.94)	2.44* (1.36)
Total	0.69 (1.83)	7.86*** (1.86)	5.53** (2.59)

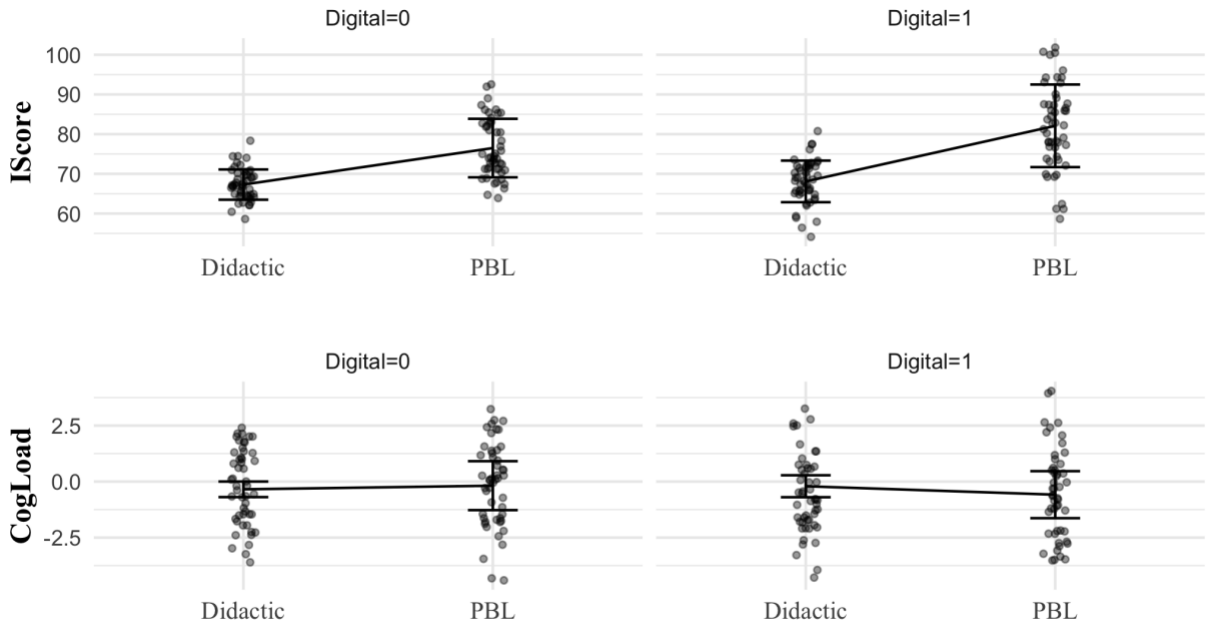
Source: Compiled by author

The results presented in Table 5.17 indicate that previous PBL experience, as represented by the variable $CumuPBL$, has no direct or indirect impact on students' immediate learning efficacy, measured by the variable $IScore$. In contrast, the positive effect of the PBL treatment, represented by the variable PBL, is solely due to its direct impact on $IScore$. However, the overall positive impact of the interaction term between PBL and $CumuPBL$ is primarily driven by its indirect impact through reducing students' cognitive load, as represented by the variable $CogLoad$. These findings suggest that students' previous PBL experience may enhance the benefits of the PBL approach by reducing cognitive load.

5.4.2 Contextual effect of digital assistant environment

Figure 5.6 presents a graphical representation of the relationship between the digital assistant environment and the efficacy of the PBL approach. The lines connecting the error bars represent the impact of the PBL treatment on students who either received or did not receive additional support from a digital assistant in Experiment 5B.

Figure 5.6: Digital assistant environment and PBL treatment effects



Source: Compiled by author

The impact of the PBL approach on students' immediate test scores ($IScore$) and cognitive load ($CogLoad$) is illustrated in Figure 5.6. This representation is based on data collected from Experiment 5B, in which students were either given or not given an additional digital learning assistant in the form of interactive simulation software Rainer.

The results depicted in Figure 5.6 suggest that the PBL treatment was effective in enhancing students' immediate test scores, regardless of whether or not they received a digital learning assistant. Additionally, the line connecting the Didactic and PBL groups for students who received the digital assistant appears to be steeper, which could indicate that the positive impact of the PBL treatment may be more pronounced in the presence of the digital assistant environment.

Regarding students' cognitive load, the results shown in Figure 5.6 suggest that the PBL treatment had a slight increase in cognitive load for students without a digital assistant, while it had a slight decrease in cognitive load for students with a digital assistant.

5.4.2.1 Two-way variance analysis

The results of the quantitative analysis examining the influence of the digital assistant environment on the efficacy of the PBL approach are presented in Table 5.18. The table displays the results of a two-way variance analysis.

Table 5.18: Two-way variance analysis for contextual effects of digital assistant environment

Variance analysis	Interaction	IScore	DScore	CogLoad
ANOVA	Two-way	5.63** (0.02)	10.76*** (0.00)	5.49** (0.02)
ANCOVA	Two-way	5.71** (0.02)	10.30*** (0.00)	5.15** (0.02)
MANOVA	Two-way	4.21*** (0.01)		
MANCOVA	Two-way	4.08*** (0.01)		

Source: Compiled by author

Table 5.18 reveals that the interaction between the PBL and $Digital$ variables has a statistically significant effect on the variance of the $IScore$, $DScore$, and $CogLoad$ outcome variables. This finding suggests that the digital assistant environment plays an important role in determining the efficacy of the PBL approach.

5.4.2.2 Linear regression analysis

The results of the linear regression analysis that aimed at examining the interaction effect of the PBL approach and the digital assistant environment on students' learning efficacy are presented in Table 5.19. The data used in this analysis was obtained from Experiment 5B.

Table 5.19: Linear regression analysis for contextual effects of digital assistant environment

Dependent Variable	IScore	DScore	CogLoad
Intercept	35.878 (24.73)	4.463 (28.75)	0.488 (2.84)
PBL	9.182*** (1.45)	10.370*** (1.68)	0.176 (0.17)

Dependent Variable	IScore	DScore	CogLoad
Digital	0.610 (1.44)	0.845 (1.67)	0.142 (0.17)
PBL & Digital	4.878** (2.04)	7.614*** (2.37)	-0.532** (0.23)
Adj. R squared	0.42	0.46	0.01
N	200	200	200

Source: Compiled by author

The model behind Table 5.19 includes the independent variables of PBL, $Digital$, and their interaction term. The coefficients, standard errors, and levels of significance, indicated by asterisks, with *, **, and *** representing significance levels of 0.1, 0.05, and 0.01 respectively, are provided in each cell. The response variables and treatment groups, as well as their definitions and constructs, can be found in Table 4.2.

The results from the linear regression analysis demonstrate that in Experiment 5B, PBL had a significant and positive correlation with both immediate test scores ($IScore$) and delayed test scores ($DScore$), in line with the findings from Experiment 1 detailed in Section 5.1. Additionally, the results indicate that PBL had no significant effect on students' cognitive load, which is in agreement with previous findings presented in Table 5.2, but contradicts the predictions of CLT. The results also demonstrate that $Digital$ was not associated with either test scores or cognitive load, suggesting that the digital assistant environment does not have a direct impact on learning outcomes in Experiment 5B.

The interaction term between PBL and $Digital$ was found to have a positive association with both immediate and delayed test scores, indicating that the benefits of PBL treatment are more pronounced in the presence of a digital assistant environment. Additionally, the interaction term was negatively correlated with cognitive load at a 5% significance level, suggesting that the enhanced benefits of PBL for students with digital assistant environment may be due, in part, to reduced cognitive load.

5.4.2.3 Path analysis

Table 5.20 presents the results of a path analysis that was conducted to assess the effect of the digital assistant environment on the effectiveness of the PBL approach.

Table 5.20: Path analysis for digital assistant environment

Path Analysis	Digital	PBL	Digital & PBL
Direct	0.98 (1.35)	9.64*** (1.36)	3.50* (1.93)
Indirect	-0.37 (0.43)	-0.46 (0.43)	1.38** (0.67)
Total	0.61 (1.41)	9.18*** (1.42)	4.88** (2.00)

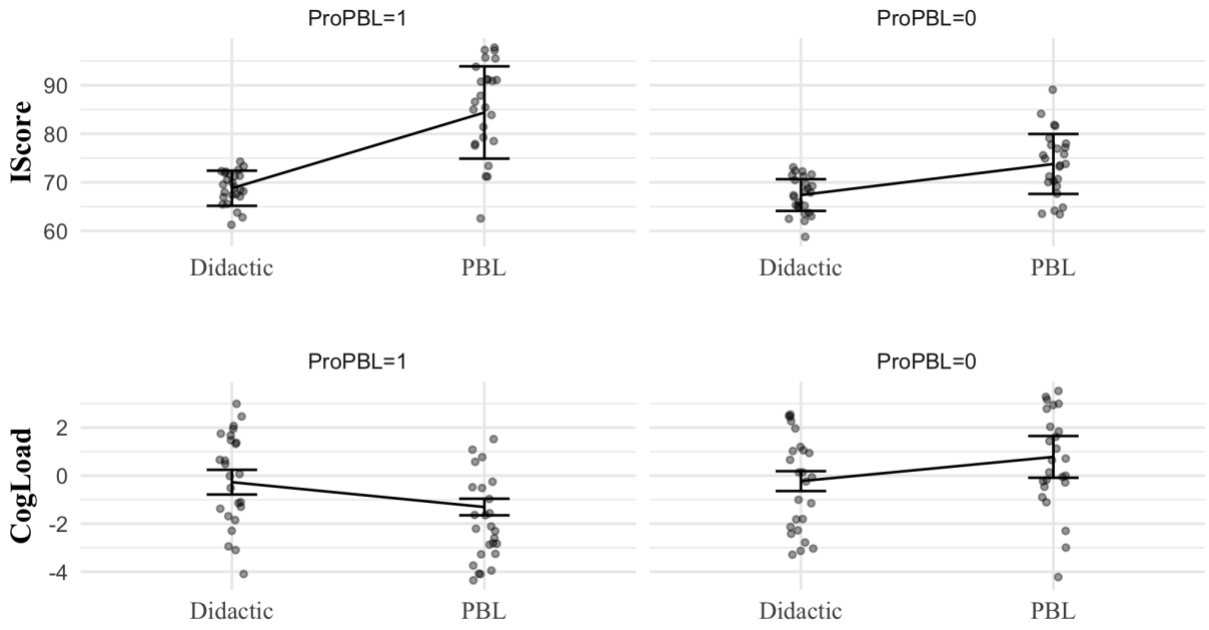
Source: Compiled by author

Table 5.20 reveals that the digital assistant environment, as represented by the variable $Digital$, does not have a direct or indirect influence on students' immediate learning efficacy, as measured by the variable $IScore$. On the other hand, the positive impact of the PBL treatment, represented by the variable PBL, is derived solely from its direct effect on $IScore$. However, the overall positive effect of the interaction between PBL and $Digital$ is the result of both its direct impact on $IScore$ and its indirect impact via a reduction in students' cognitive load, represented by the variable $CogLoad$. These results suggest that the digital assistant environment may enhance the benefits of PBL by reducing cognitive load as well as other factors beyond it.

5.4.3 Contextual effect of pro-PBL family culture

The relationship between the pro-PBL family culture and the efficacy of the PBL approach is presented graphically in Figure 5.7. The figure illustrates the impact of the PBL treatment on students who are from a pro-PBL family and those who are not, as indicated by the connecting lines between the error bars in Experiment 6.

Figure 5.7: Pro-PBL family culture and PBL treatment effects



Source: Compiled by author

The graphical representation in Figure 5.7 provides evidence for the efficacy of the PBL treatment in enhancing students' immediate test scores, regardless of the presence or absence of a pro-PBL family culture. In this study, a pro-PBL family culture is defined as a family where the parent has an overseas educational background; it is represented by the variable $ProPBL = 1$ (as further defined in Table 4.2). Furthermore, the steepness of the line connecting the Didactic and PBL groups for students from pro-PBL families appears to suggest that the positive impact of the PBL treatment is more pronounced in the presence of such a culture. The results depicted in Figure 5.7 also indicate that the PBL treatment leads to an increase in cognitive load for students from non-pro-PBL families, while it results in a reduction of cognitive load for students from pro-PBL families.

5.4.3.1 Two-way variance analysis

The findings of the quantitative examination of the impact of pro-PBL family culture on the effectiveness of the PBL approach can be found in Table 5.21. The results are displayed in the form of a two-way variance analysis between PBL and $ProPBL$.

Table 5.21: Two-way variance analysis for contextual effects of pro-PBL family culture

Variance analysis	Interaction	IScore	DScore	CogLoad
ANOVA	Two-way	13.91*** (0.00)	14.94*** (0.00)	79.63*** (0.00)
ANCOVA	Two-way	13.02*** (0.00)	12.77*** (0.00)	72.88*** (0.00)
MANOVA	Two-way	26.44*** (0.00)		
MANCOVA	Two-way	24.23*** (0.00)		

Source: Compiled by author

The findings in Table 5.21 indicate that the interaction between the PBL and $ProPBL$ variables has a statistically significant impact on the variances of the $IScore$, $DScore$, and $CogLoad$ outcome variables. This result implies that pro-PBL family culture is important in influencing the effectiveness of PBL.

5.4.3.2 Linear regression analysis

The findings of a linear regression analysis aimed at evaluating the impact of the interaction between the PBL and $ProPBL$ variables on students' learning efficacy, as measured by the $IScore$, $DScore$, and $CogLoad$ outcome variables, are presented in Table 5.22. These results were derived from the data collected during Experiment 6.

Table 5.22: Linear regression analysis for contextual effects of the pro-PBL family culture

Dependent Variable	IScore	DScore	CogLoad
Intercept	32.232 (31.50)	72.800** (34.53)	-2.232 (2.95)
PBL	6.613*** (1.75)	4.698** (1.92)	0.989*** (0.16)
ProPBL	1.598 (1.75)	0.654 (1.91)	-0.053 (0.16)
PBL & ProPBL	8.997*** (2.49)	9.767*** (2.73)	-1.996*** (0.23)
Adj. R squared	0.54	0.45	0.61

Dependent Variable	IScore	DScore	CogLoad
N	100	100	100

Source: Compiled by author

The model behind Table 5.22 includes the independent variables of PBL, $ProPBL$, and their interaction term. The coefficients, standard errors, and levels of significance, as indicated by asterisks, with * representing 0.1, ** representing 0.05, and *** representing 0.01 significance levels, are provided in the table.

The results of the linear regression analysis reveal that in Experiment 6, the PBL approach has a significant and positive correlation with both immediate test scores ($IScore$) and delayed test scores ($DScore$), which is in line with the findings from Experiment 1 presented in Section 5.1. Furthermore, the results indicate that the PBL approach has a significant positive effect on students' cognitive load, which is consistent with the predictions of CLT but not observed in other experiments included in this thesis. Additionally, the results demonstrate that $ProPBL$ does not have a direct impact on the learning outcomes in Experiment 6, as it was not associated with either test scores or cognitive load.

The interaction term between the PBL approach and pro-PBL family culture was found to have a positive association with both immediate and delayed test scores, indicating that the benefits of the PBL treatment are more pronounced in students who come from a pro-PBL family culture. Additionally, the interaction term was negatively correlated with cognitive load at a 1% significance level, which suggests that the enhanced benefits of the PBL approach for students from a pro-PBL family culture may be due to reduced cognitive load.

5.4.3.3 Path analysis

The results of a path analysis that aimed to evaluate the impact of the pro-PBL family culture on the efficacy of the PBL approach are presented in Table 5.23.

Table 5.23: Path analysis for the pro-PBL family culture

Path Analysis	ProPBL	PBL	ProPBL & PBL
Direct	1.42 (1.59)	9.98*** (1.88)	2.21 (3.04)

Path Analysis	ProPBL	PBL	ProPBL & PBL
Indirect	0.18 (0.54)	-3.36*** (1.13)	6.78*** (2.16)
Total	1.60 (1.68)	6.61*** (1.68)	9.00*** (2.39)

Source: Compiled by author

Table 5.23 provides evidence that the pro-PBL family culture, as represented by the variable $ProPBL$, does not have a direct or indirect effect on the immediate learning efficacy of students, as measured by the variable $IScore$. In contrast, the positive impact of the PBL treatment, represented by the variable PBL, can be attributed to a combination of direct and indirect effects. The direct effect of PBL on $IScore$ is positive, indicating an improvement in students' learning outcomes beyond the cognitive load channel. However, the indirect effect of PBL on $IScore$ is negative, resulting from increased cognitive load. Despite this, the overall effect of PBL on $IScore$ is positive, as its direct effect is more pronounced. For the interaction between PBL and $ProPBL$, the overall positive effect on $IScore$ is driven primarily by the reduction of cognitive load, as the indirect effect is significant, while the direct effect is not. In conclusion, the results suggest that the pro-PBL family culture can enhance the efficacy of the PBL approach by alleviating students' cognitive load and preserving their working memory.

5.5 Summary of Chapter 5

Chapter 5 presents the principal empirical findings derived from the research design and sample described in Chapter 3. To facilitate a clear understanding of the results, a comparison between the findings and the hypotheses posited in Chapter 3 will be made. The comparison for RQ1 and RQ2 is presented in Table 5.24, which mirrors the structure of Table 3.2 but includes additional columns indicating the empirical findings for each prediction.

Table 5.24: Comparison of empirical findings with predictions (RQ1 & RQ2, 'P' for prediction and 'F' for finding)

	Moderate guidance treatment		Minimal guidance treatment		Late guidance treatment		Collaborative learning treatment		Applied PBL approach	
	P	**F**	**P**	**F**	**P**	**F**	**P**	**F**	**P**	**F**
Cognitive load score	?	+	?	+	?	+	?	?	?	?
Knowledge acquisition score	+	?	-	-	+	?	+	+	+	+
Enjoyment score	+	?	+	?	+	?	+	+	+	+
Self-efficacy score	+	+	+	?	+	+	+	+	+	+

Source: Compiled by author

The empirical findings of the research in this thesis are largely consistent with the predictions posited in Chapter 3. The PBL approach was consistently found to improve students' learning outcomes in all six experiments, as reflected in higher testing scores, increased enjoyment in learning, and enhanced self-efficacy with regards to the knowledge gained. The predictions of CLT, which argues that PBL would result in increased cognitive load and depleted working memory, were partially supported by the *post hoc* results of Experiment 1. However, the results of the other contextual experiments were not in line with this proposition, with the exception of Experiment 6. Even in Experiment 6, the positive effects of PBL on learning outcomes through factors beyond cognitive load outweighed any negative impacts on learning through cognitive load. The results from the majority of the path analyses (Experiments 3 to 5) indicate that the effects of PBL on learning outcomes primarily stem from mechanisms beyond cognitive load.

The empirical findings with regards to the pedagogical elements of PBL indicate that moderate and minimal levels of guidance, as well as late guidance, resulted in increased cognitive load, as evidenced in Experiment 1. The impact of collective learning on cognitive load was not clear. The testing results did not reveal a clear impact on testing scores for moderate guidance and late guidance, but minimal guidance was found to reduce the immediate testing score by 8.5 marks and the delayed testing score by 12.8 marks compared to traditional didactic teaching methods, as documented in Table 5.2. While collective learning alone improved the delayed testing scores by 3.3 marks, this improvement was not significant at the 5% level. All PBL elements were found to improve students' self-efficacy, while only collective learning had a positive impact on students' enjoyment, as documented in the testing results. These findings confirm the inverted U-

shaped relationship between guidance level and learning efficacy, as well as the unique benefit of collective learning compared to other elements.

The empirical findings regarding the effect of the PBL approach suggest a positive synergistic effect on short-term learning outcomes, as evidenced by the marginally significant results. Table 5.3 indicates that two-way and three-way classifications contribute to explaining the variance of the short-term learning outcome indicators, while Table 5.4 shows that the three-way interaction term has a positive impact on testing scores at a 10% significance level. The positive synergistic effect of the PBL approach was observed for students' enjoyment in the long-term learning outcome indicators, but the impact on students' self-efficacy was unclear, as documented in Tables 5.7 and 5.8.

The comparison between the predictions outlined in Chapter 3 and the empirical findings for RQ3 and RQ4 is depicted in Table 5.25, which parallels the format of Table 3.3 but incorporates additional columns exhibiting the actual results for each prediction.

Table 5.25: Comparison of empirical findings with predictions (RQ3 & RQ4)

Contextual variable	Response variable	Prediction	Finding
Higher-grade students	Cognitive load score	?	?
	Knowledge acquisition score	+	+
Knowledge with more interactive elements	Cognitive load score	?	?
	Knowledge acquisition score	-	-
PBL for one semester	Cognitive load score	-	-
	Knowledge acquisition score	+	+
PBL with more digital assistance	Cognitive load score	-	-
	Knowledge acquisition score	+	+
Family with pro-PBL cultural background	Cognitive load score	-	-
	Knowledge acquisition score	+	+

Source: Compiled by author

As is evident from Table 3.3, this thesis documents results consistent with the predictions in Chapter 3 for the two contextual factors suggested by CLT and modified CLTs. Specifically, the interaction term of $Hgrade$ and PBL was found to have a significantly positive impact on

$IScore$, as illustrated in Table 5.10. Additionally, the interaction term of $Complex$ and PBL was shown to have a significantly positive impact on both $IScore$ and $DScore$, as presented in Table 5.13. Although these results are consistent with the predictions of CLT with regards to knowledge acquisition score, the underlying logic revealed by this thesis is different from that predicted by CLT. The higher grade students did not experience a decrease in cognitive load when receiving PBL treatment, and the more complex learning task did not result in an increased cognitive load for the PBL treatment groups. The path analyses in Tables 5.11 and 5.14 suggest that the impact of these interaction terms is mainly through direct channels, rather than indirect channels related to cognitive load. Thus, the findings of this thesis suggest that prior knowledge and task complexity could influence the efficacy of PBL, but not in accordance with the logic proposed by CLT.

This thesis also demonstrates results consistent with the predictions for the three contextual factors relevant to RQ4. The results in Tables 5.19 and 5.19 reveal that previous PBL-like learning experiences can help reduce cognitive load in students receiving PBL treatment, which leads to improved testing scores. Additionally, digital assistant environments were found to significantly reduce cognitive load in PBL treatment groups, leading to a more pronounced PBL benefit with respect to testing scores. Furthermore, pro-PBL family cultures were found to reduce cognitive load for students in PBL treatment groups, resulting in even higher scores, as evidenced in Table 5.22. The path analyses in Tables 5.17, 5.20, and 5.23 suggest that the contextual effects of previous PBL and pro-PBL family culture are dominated by indirect effects through cognitive load, while digital assistance environments have both direct and indirect effects. These empirical answers to RQ4 suggest that the CLT framework should incorporate additional contextual factors, as advised by the ICLT.

6 Conclusions

This thesis aimed to address two research gaps in the field of PBL by raising four research questions and conducting six experiments in two schools in China over a period of two years and nine months. The analyses of this thesis are based on a total sample size of 2,334 students from grades 8 and 9. The findings demonstrate the positive efficacy of PBL on students' learning outcomes and provide evidence against some of the arguments of CLT. This thesis also explored the effect of individual pedagogical elements of PBL and revealed the advantage of collective learning and the positive synergistic effect of a PBL approach. Furthermore, the findings show that prior knowledge, task complexity, previous PBL-like learning experiences, digital assistance environment, and pro-PBL family culture can all influence the efficacy of PBL. These findings suggest that the CLT framework should incorporate additional contextual factors, as advised by modified CLTs such as ICLT.

In this thesis, several empirical design strategies have been employed to ensure the validity and reliability of the results. First, a stratified random sampling strategy was used to minimize disparities across the experimental groups. By employing this method, I was able to account for systematic differences between the treatment and control groups, such as pre-treatment testing scores, that may impact student learning outcomes. Second, a pre-recorded video was utilized to isolate the treatment variable and control for other factors that might influence the learning environment. This method helped to standardize all other aspects of the learning environment, such as the teacher, teaching material, class duration, and class size, to minimize any discrepancies that may arise as a result of differences among teachers. Third, path analysis was conducted to distinguish direct and indirect effects of PBL and its contextual factors. Path analysis provides a means to differentiate between PBL's direct impact on learning outcomes and its indirect impact through cognitive load.

This thesis makes several contributions to the field of PBL research. First, it offers empirical evidence from a non-western cultural context, through a series of randomized controlled experiments in China. These experiments deepen our understanding of the impact of PBL on cognitive load and learning outcomes in this particular cultural setting. Second, this thesis goes beyond a holistic view of PBL by examining the individual effects of its elements, and provides

evidence of the positive synergistic efficacy of the approach. This aspect has been underrepresented in previous studies. Third, this thesis identifies and provides evidence of additional contextual factors influencing the efficacy of PBL, including the family-level culture, prior PBL experience, and digital assistance. This contributes to a more comprehensive understanding of the effectiveness of PBL in a non-western cultural context and the factors that impact its efficacy.

6.1 Limitations

However, it is important to acknowledge the limitations of this thesis that should be taken into account when interpreting the results. First, the generalizability of the findings may be limited. The experiments were conducted in only two secondary schools in China, specifically focusing on the teaching of physics. The results may not be applicable to other countries, regions, grade levels, disciplines, or educational institutions, thereby limiting the representation of the findings.

Second, the thesis may be subject to measurement error. For instance, it is questionable whether student enjoyment and self-efficacy serve as appropriate measures for long-term learning outcomes. Although it is plausible that students who enjoy learning and have confidence in their knowledge may have better future learning outcomes, there is still potential for measurement error. Other variables, such as critical thinking skills and future career concerns, should also be considered. For treatment variables, utilizing various combinations of pre-recorded instructional videos may result in a discrepancy from the 'naturalistic' settings of PBL teaching. This difference imposes a limitation on drawing inferences from the results of this study. For contextual variables, the extent to which a parent’s overseas education background can represent a family-level pro-PBL culture is also subject to measurement error and would benefit from more direct measures. Additionally, the measurements of 'digital assistance' and 'pro-PBL' as binary variables may oversimplify the intricate processes involved in teaching using digital technologies and the influence of family background.

Third, the issue of causality is not fully addressed in this thesis. The causality between the PBL treatment and the response variable depends on the randomness of the sampling. Despite the use of stratified random sampling to mitigate disparities across experimental groups, there may still be unobserved differences between groups that contribute to differences in learning outcomes.

Thus, it is in principle possible that it is not the treatment group that leads to the pattern of learning efficacy, but rather these unobserved differences.

6.2 Recommendations

This section offer several directions for future research based on the implications and limitations of this thesis. First, it would be valuable to expand the scope of research on PBL in non-western cultural contexts. This would provide a more comprehensive understanding of the utilization of PBL in different countries, regions, grade levels, disciplines, and educational institutions. By conducting studies in multiple non-western cultures, we can overcome the limitation of representation in this thesis.

Second, further exploration into the individual elements and synergistic effects of PBL is of great interest. This thesis has demonstrated the significance of collective learning; however, more research is needed to understand why collective learning is more beneficial. Investigating the synergistic effects of the PBL approach across different countries, including western countries, would deepen our understanding of the optimal combination of PBL elements.

Third, the measurement of long-term learning outcomes is a crucial area for future PBL-related research. More studies should focus on this aspect and explore alternative variables, such as critical thinking skills and future career concerns. It is also necessary to examine the correlation between these variables and their underlying latent factors. In addition, a longer window sample, such as a 10-year period, could be used to evaluate which factors best predict students' actual future scientific achievements.

Fourth, the family-level cultural factor is a fascinating area for further investigation. More studies that use direct measures, such as standardized questionnaires or assessments of parents' attitudes towards PBL, would provide a more nuanced understanding of the family's cultural status regarding PBL. Micro data on the family's parenting history could also provide insights into how family culture is shaped and inform future research in this area.

In addition to the academic community, this thesis also holds implications for various stakeholders in education, including teachers and policy makers. Teachers are encouraged to incorporate more digital scaffolding in conjunction with PBL teaching, as the thesis indicates

that this combination can significantly enhance the efficacy of PBL. Schools should also aim for a sustained, long-term implementation of PBL, as students' increased familiarity with this method can lead to more efficient learning. For education policy makers in non-western cultures, such as China, this research provides evidence that PBL is an effective strategy for secondary school education, despite the distinct social norms compared to western societies. This study also enlightens policy makers about the importance of modifying societal cultural attitudes to be more supportive of the PBL approach, in order to maximize its impact in schools.

As I embarked on the first experiment of this thesis in 2016, I could not have foreseen the profound impact technology would have on the world and the learning environment by the time my thesis was completed. The past three years of global pandemic have seen the rapid evolution of artificial intelligence, exemplified by the emergence of transformative AI products like ChatGPT. The way we learn and acquire knowledge has undergone dramatic changes throughout history, and we can only imagine what the students of the Victorian era would think of our present-day methods. However, the exponential growth of AI and other technological innovations has brought forth a pressing concern as to what the future students should be equipped with, now more imperative than ever before. A growing body of research (García-Peñalvo, 2023; Pavlik, 2023; Zhai, 2022) has explored how current learning methods are being affected by new AI products like ChatGPT. It is against this backdrop that my thesis is being presented, at a time when the importance of evaluating the effectiveness of various learning methods, including the PBL approach, cannot be overstated.

References

Abdi, H., and Williams, L. J. (2010). Tukey's honestly significant difference (HSD) test. *Encyclopedia of Research Design*, *3*(1), 1–5.

Aditomo, A., and Klieme, E. (2020). Forms of inquiry-based science instruction and their relations with learning outcomes: Evidence from high and low-performing education systems. *International Journal of Science Education*, *42*(4), 504–525.

Aidoo, B., Boateng, S. K., Kissi, P. S., and Ofori, I. (2016). Effect of Problem-Based Learning on Students' Achievement in Chemistry. *Journal of Education and Practice*, *7*(33), 103–108.

Ajai, J. T., Imoko, B. I., and O'kwu, E. I. (2013). Comparison of the learning effectiveness of problem-based learning (PBL) and conventional method of teaching algebra. *Journal of Education and Practice*, *4*(1), 131–135.

Alfieri, L., Brooks, P. J., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1.

Allen, D. K., Brown, A., Karanasios, S., and Norman, A. (2013). How should technology-mediated organizational change be explained? A comparison of the contributions of critical realism and activity theory. *Mis Quarterly*, 835–854.

Alloway, T. P., Moulder, R., Horton, J. C., Leedy, A., Archibald, L. M., Burin, D., Injoque-Ricle, I., Passolunghi, M. C., and Dos Santos, F. H. (2017). Is it a small world after all? Investigating the theoretical structure of working memory cross-nationally. *Journal of Cognition and Culture*, *17*(3-4), 331–353.

Amalia, E., Surya, E., and Syahputra, E. (2017). The effectiveness of using problem based learning (PBL) in mathematics problem solving ability for junior high school students. *International Journal of Advance Research and Innovative Ideas in Education*, *3*(2), 3402–3406.

Areepattamannil, S. (2012). Effects of inquiry-based science instruction on science achievement and interest in science: Evidence from Qatar. *The Journal of Educational Research*, *105*(2), 134–146.

Argaw, A. S., Haile, B. B., Ayalew, B. T., and Kuma, S. G. (2016). The effect of problem based learning (PBL) instruction on students' motivation and problem solving skills of physics. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(3), 857–871.

Ashman, G. (2022). *Should problem solving precede explicit instruction when element interactivity is high?* [Doctoral dissertation]. UNSW Sydney.

Ashman, G., Kalyuga, S., and Sweller, J. (2020). Problem-solving or explicit instruction: Which should go first when element interactivity is high? *Educational Psychology Review*, *32*(1), 229–247.

Ayres, P. (2017). Subjective measures of cognitive load: What can they reliably measure? In *Cognitive Load Measurement and Application* (pp. 9–28). Routledge.

Baddeley, A. (2012). How does emotion influence working memory. In *Attention, representation, and human performance: Integration of cognition, emotion, and motivation* (pp. 3–18). Psychology Press.

Bai, Y. (2019). Farewell to confucianism: The modernizing effect of dismantling China's imperial examination system. *Journal of Development Economics*, *141*, 102382.

Barth, V. L., Piwowar, V., Kumschick, I. R., Ophardt, D., and Thiel, F. (2019). The impact of direct instruction in a problem-based learning setting. Effects of a video-based training program to foster preservice teachers' professional vision of critical incidents in the classroom. *International Journal of Educational Research*, *95*, 1–12.

Blayney, P., Kalyuga, S., and Sweller, J. (2016). The impact of complexity on the expertise reversal effect: Experimental evidence from testing accounting students. *Educational Psychology*, *36*(10), 1868–1885.

Brünken, R. E., Plass, J. L., and Moreno, R. E. (2010). Current issues and open questions in cognitive load research. In *Cognitive Load Theory* (pp. 253–272). Cambridge University Press.

Cairns, D., and Areepattamannil, S. (2019). Exploring the relations of inquiry-based teaching to science achievement and dispositions in 54 countries. *Research in Science Education*, *49*(1), 1–23.

Cebolla-Boado, H., Hu, Y., and Soysal, Y. N. (2018). Why study abroad? Sorting of Chinese students across British universities. *British Journal of Sociology of Education*, *39*(3), 365–380.

Chase, C. C., and Klahr, D. (2017). Invention versus direct instruction: For some content, it'sa tie. *Journal of Science Education and Technology*, *26*(6), 582–596.

Chen, O., and Kalyuga, S. (2020). Exploring factors influencing the effectiveness of explicit instruction first and problem-solving first approaches. *European Journal of Psychology of Education*, *35*(3), 607–624.

Chen, O., and Kalyuga, S. (2021). Working Memory Resources Depletion Makes Delayed Testing Beneficial. *Journal of Cognitive Education and Psychology*, *20*(1), 38–46.

Chen, O., Kalyuga, S., and Sweller, J. (2016). When instructional guidance is needed. *The Educational and Developmental Psychologist*, *33*(2), 149–162.

Chen, O., Kalyuga, S., and Sweller, J. (2017). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review*, *29*(2), 393–405.

Chen, O., Retnowati, E., and Kalyuga, S. (2019). Effects of worked examples on step performance in solving complex problems. *Educational Psychology*, *39*(2), 188–202.

Chen, O., Retnowati, E., and Kalyuga, S. (2020). Element interactivity as a factor influencing the effectiveness of worked example–problem solving and problem solving–worked example sequences. *British Journal of Educational Psychology*, *90*, 210–223.

Chen, O., Woolcott, G., and Kalyuga, S. (2021). Comparing alternative sequences of examples and problem-solving tasks: The case of conceptual knowledge. *The Educational and Developmental Psychologist*, *38*(1), 158–170.

Chen, T., Kung, J. K., and Ma, C. (2020). Long live Keju! The persistent effects of China's civil examination system. *The Economic Journal*, *130*(631), 2030–2064.

Chen, Y., Huang, R., Lu, Y., and Zhang, K. (2021). Education fever in China: Children's academic performance and parents' life satisfaction. *Journal of Happiness Studies*, *22*(2), 927–954.

Chin, D. B., Chi, M., and Schwartz, D. L. (2016). A comparison of two methods of active learning in physics: Inventing a general solution versus compare and contrast. *Instructional Science*, *44*(2), 177–195.

Choon-Eng Gwee, M. (2008). Globalization of problem-based learning (PBL): Cross-cultural implications. *The Kaohsiung Journal of Medical Sciences*, *24*, S14–S22.

Clark, R. E. (2009). How much and what type of guidance is optimal for learning from instruction? In *Constructivist instruction* (pp. 170–195). Routledge.

Cochran, W. G. (2011). Sampling techniques. In *Sampling techniques* (pp. 428–428).

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

Cohen, L., Manion, L., and Morrison, K. (2002). *Research methods in education*. routledge.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Dabbagh, N. (2019). Effects of PBL on critical thinking skills. *The Wiley Handbook of Problem-Based Learning*, 135–156.

Dagyar, M., and Demirel, M. (2015). Effects of problem-based learning on academic achievement: A meta-analysis study. *Education and Science*, *40*(181), 139–174.

Darabi, A., Arrington, T. L., and Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development*, *66*(5), 1101–1118.

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., and Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, *24*(2), 97–140.

Dawson, J. F., and Richter, A. W. (2006). Probing three-way interactions in moderated multiple regression: Development and application of a slope difference test. *Journal of Applied Psychology*, *91*(4), 917.

De Witte, K., and Rogge, N. (2016). Problem-based learning in secondary education: Evaluation by an experiment. *Education Economics*, *24*(1), 58–82.

Debue, N., and Van De Leemput, C. (2014). What does germane load mean? An empirical contribution to the cognitive load theory. *Frontiers in Psychology*, *5*, 1099.

DeLeeuw, K. E., and Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, *100*(1), 223.

Dello-Iacovo, B. (2009). Curriculum reform and 'quality education'in China: An overview. *International Journal of Educational Development*, *29*(3), 241–249.

Deudney, D., and Ikenberry, G. J. (2009). The myth of the autocratic revival: Why liberal democracy will prevail. *Foreign Affairs*, 77–93.

Dolmans, D. H., Loyens, S. M., Marcq, H., and Gijbels, D. (2016). Deep and surface learning in problem-based learning: A review of the literature. *Advances in Health Sciences Education*, *21*(5), 1087–1112.

Du, X., and Chaaban, Y. (2020). Teachers' Readiness for a Statewide Change to PjBL in Primary Education in Qatar. *Interdisciplinary Journal of Problem-Based Learning*, *14*(1), n1.

Duchi, L., Lombardi, D., Paas, F., and Loyens, S. M. (2020). How a growth mindset can change the climate: The power of implicit beliefs in influencing people's view and action. *Journal of Environmental Psychology*, *70*, 101461.

Dunn, T. J., Baguley, T., and Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*(3), 399–412.

Education, C. M. of. (2020). *Statistics on Chinese learners studying overseas in 2019*. http://en.moe.gov.cn/news/press_releases/202012/t20201224_507474.html

Education, C. M. of. (2001). *Science Curriculum Standards (7–9 Grades) of Full-Time Compulsory Education (Trial Version)*. Beijing Normal University Press. http://edu.cn/20010926/3002911.shtml

Education, C. M. of. (2002). *The Concept of Quality Education: Key Points for Study*. Beijing Normal University Press.

Education, C. M. of. (2011). *National Science Curriculum Standards (7–9 Grades) of the Compulsory Education*. Beijing Normal University Press.

English, M., and Kitsantas, A. (2019). PBL Capstone Experience in Conservation Biology: A Self-Regulated Learning Approach. *The Wiley Handbook of Problem-Based Learning*, 507–527.

Ertmer, P. A., and Glazewski, K. D. (2015). Essentials for PBL implementation: Fostering collaboration, transforming roles, and scaffolding learning. *Essential Readings in Problem-Based Learning*, *58*, 89–106.

Ertmer, P. A., and Glazewski, K. D. (2019). Scaffolding in PBL environments: Structuring and problematizing relevant task features. *The Wiley Handbook of Problem-Based Learning*, 321–342.

Feldon, D. F., Callan, G., Juth, S., and Jeong, S. (2019). Cognitive load as motivational cost. *Educational Psychology Review*, *31*(2), 319–337.

Firdaus, F. M., and Herman, T. (2017). Improving Primary Students' Mathematical Literacy through Problem Based Learning and Direct Instruction. *Educational Research and Reviews*, *12*(4), 212–219.

Fonteijn, H. T., and Dolmans, D. H. (2019). Group work and group dynamics in PBL. *The Wiley Handbook of Problem-Based Learning*, 199–220.

Forbes, C. T., Neumann, K., and Schiepe-Tiska, A. (2020). Patterns of inquiry-based science instruction and student science achievement in PISA 2015. *International Journal of Science Education*, *42*(5), 783–806.

Fraenkel, J. R., Wallen, N. E., and Hyun, H. H. (2012). *How to design and evaluate research in education* (Vol. 7). McGraw-hill New York.

Frambach, J. M., Driessen, E. W., Chan, L.-C., and Vleuten, C. P. van der. (2012). Rethinking the globalisation of problem-based learning: How culture challenges self-directed learning. *Medical Education*, *46*(8), 738–747.

Fraser, K. L., Ayres, P., and Sweller, J. (2015). Cognitive load theory for the design of medical simulations. *Simulation in Healthcare*, *10*(5), 295–307.

Frederiksen, C. H. (2018). Learning to reason through discourse in a problem-based learning group. In *Discourse Processes* (pp. 135–160). Routledge.

Fujikoshi, Y. (1993). Two-way ANOVA models with unbalanced data. *Discrete Mathematics*, *116*(1-3), 315–334.

Furtak, E. M., Seidel, T., Iverson, H., and Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research*, *82*(3), 300–329.

Galecki, A., Burzykowski, T., Ga\lecki, A., and Burzykowski, T. (2013). *Linear mixed-effects model*. Springer.

Galy, E., Cariou, M., and Mélan, C. (2012). What is the relationship between mental workload factors and cognitive load types? *International Journal of Psychophysiology*, *83*(3), 269–275.

Gao, S., and Wang, J. (2014). Teaching transformation under centralized curriculum and teacher learning community: Two Chinese chemistry teachers' experiences in developing inquiry-based instruction. *Teaching and Teacher Education*, *44*, 1–11.

Gao, S., Wang, J., and Zhong, Z. (2018). Influence of science instruction reform on academic performance of eighth grade students in Chinese inner-Mongolia autonomous region. *Compare: A Journal of Comparative and International Education*, *48*(6), 879–895.

Gao, X., Wang, L., Deng, J., Wan, C., and Mu, D. (2022). The effect of the problem based learning teaching model combined with mind mapping on nursing teaching: A meta-analysis. *Nurse Education Today*, *111*, 105306.

García-Peñalvo, F. J. (2023). *The perception of Artificial Intelligence in educational contexts after the launch of ChatGPT: Disruption or Panic?*

Geary, D. C. (2008). An evolutionarily informed education science. *Educational Psychologist*, *43*(4), 179–195.

Geary, D. C. (2012). *Evolutionary educational psychology.*

Geary, D. C., and Berch, D. B. (2016). Evolution and children's cognitive and academic development. In *Evolutionary perspectives on child development and education* (pp. 217–249). Springer.

Gillies, R. M. (2016). Cooperative learning: Review of research and practice. *Australian Journal of Teacher Education (Online)*, *41*(3), 39–54.

Grange, J. (2004). *John Dewey, Confucius, and global philosophy*. SUNY Press.

Grant, M. M., and Tamim, S. R. (2019). PBL in K–12 Education. *The Wiley Handbook of Problem-Based Learning*, 221–243.

Guan, Q., and Meng, W. (2007). China's new national curriculum reform: Innovation, challenges and strategies. *Frontiers of Education in China*, *2*(4), 579–604.

Guo, L., Huang, J., and Zhang, Y. (2019). Education development in China: Education return, quality, and equity. *Sustainability*, *11*(13), 3750.

Hendarwati, E., Nurlaela, L., Bachri, B., and Sa'ida, N. (2021). Collaborative Problem based learning integrated with online learning. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(13), 29–39.

Hendriana, H., Johanto, T., and Sumarmo, U. (2018). The Role of Problem-Based Learning to Improve Students' Mathematical Problem-Solving Ability and Self Confidence. *Journal on Mathematics Education*, *9*(2), 291–300.

Hmelo-Silver, C. E., Bridges, S. M., and McKeown, J. M. (2019). Facilitating problem-based learning. *The Wiley Handbook of Problem-Based Learning*, 297–319.

Hmelo-Silver, C. E., Duncan, R. G., and Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and. *Educational Psychologist*, *42*(2), 99–107.

Holen, A., Manandhar, K., Pant, D. S., Karmacharya, B. M., Olson, L. M., and Koju, R. (2015). Medical students' preferences for problem-based learning in relation to culture and personality: A multicultural study. *International Journal of Medical Education*, *6*, 84.

Holmes, N. G., Day, J., Park, A. H., Bonn, D. A., and Roll, I. (2014). Making the failure more productive: Scaffolding the invention process to improve inquiry behaviors and outcomes in invention activities. *Instructional Science*, *42*(4), 523–538.

Hsu, C.-Y., Kalyuga, S., and Sweller, J. (2015). When should guidance be presented in physics instruction? *Archives of Scientific Psychology*, *3*(1), 37.

Hung, W., Dolmans, D. H., and Van Merriënboer, J. J. (2019). A review to identify key perspectives in PBL meta-analyses and reviews: Trends, gaps and future research directions. *Advances in Health Sciences Education*, *24*(5), 943–957.

Hung, W., Jonassen, D. H., and Liu, R. (2008). Problem-based learning. In *Handbook of research on educational communications and technology* (pp. 485–506). Routledge.

Hung, W., Moallem, M., and Dabbagh, N. (2019). Social foundations of problem-based learning. *The Wiley Handbook of Problem-Based Learning*, 51–79.

Hwang, J., Choi, K. M., Bae, Y., and Shin, D. H. (2018). Do teachers' instructional practices moderate equity in mathematical and scientific literacy?: An investigation of the PISA 2012 and 2015. *International Journal of Science and Mathematics Education*, *16*(1), 25–45.

Imafuku, R., Kataoka, R., Mayahara, M., Suzuki, H., and Saiki, T. (2014). Students' experiences in interdisciplinary problem-based learning: A discourse analysis of group interaction. *Interdisciplinary Journal of Problem-Based Learning*, *8*(2), 1.

Imanieh, M. H., Dehghani, S. M., Sobhani, A. R., and Haghighat, M. (2014). Evaluation of problem-based learning in medical students' education. *Journal of Advances in Medical Education & Professionalism*, *2*(1), 1.

Jabarullah, N. H., and Hussain, H. I. (2019). The effectiveness of problem-based learning in technical and vocational education in Malaysia. *Education+ Training*, *61*(5), 552–567.

Jerrim, J., Oliver, M., and Sims, S. (2020). The relationship between inquiry-based teaching and students' achievement. New evidence from a longitudinal PISA study in England. *Learning and Instruction*, 101310.

Jiang, D., and Kalyuga, S. (2020). Confirmatory factor analysis of cognitive load ratings supports a two-factor model. *Tutorials in Quantitative Methods for Psychology*, *16*, 216–225.

Jiang, F., and McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and attitudes: Evidence from propensity score analysis of PISA data. *International Journal of Science Education*, *37*(3), 554–576.

Jonassen, D. H. (2009). Reconciling a human cognitive architecture. In *Constructivist instruction* (pp. 25–45). Routledge.

Jonassen, D. H. (2011). Supporting problem solving in PBL. *Interdisciplinary Journal of Problem-Based Learning*, *5*(2), 95–119.

Jonassen, D. H., and Hung, W. (2015). All problems are not equal: Implications for problem-based learning. *Essential Readings in Problem-Based Learning*, 17–42.

Juandi, D., and Tamur, M. (2021). Review of problem-based learning trends in 2010-2020: A meta-analysis study of the effect of problem-based learning in enhancing mathematical problem-solving skills of Indonesian students. *Journal of Physics: Conference Series*, *1722*, 012103.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review*, *23*(1), 1–19.

Kalyuga, S., and Singh, A.-M. (2016). Rethinking the boundaries of cognitive load theory in complex learning. *Educational Psychology Review*, *28*(4), 831–852.

Kang, J., and Keinonen, T. (2018). The effect of student-centered approaches on students' interest and achievement in science: Relevant topic-based, open and guided inquiry-based, and discussion-based approaches. *Research in Science Education*, *48*(4), 865–885.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science*, *40*(4), 651–672.

Kapur, M. (2014). Productive failure in learning math. *Cognitive Science*, *38*(5), 1008–1022.

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*(2), 289–299.

Karademir, A., and Akman, B. (2019). Effect of Inquiry-Based Mathematics Activities on Preschoolers' Math Skills. *International Journal of Progressive Education*, *15*(5), 198–215.

Kaya, S., and Rice, D. C. (2010). Multilevel effects of student and classroom factors on elementary science achievement in five countries. *International Journal of Science Education*, *32*(10), 1337–1363.

Kazemi, F., and Ghoraishi, M. (2012). Comparison of problem-based learning approach and traditional teaching on attitude, misconceptions and mathematics performance of University Students. *Procedia-Social and Behavioral Sciences*, *46*, 3852–3856.

Kelly, G. J. (2014). Inquiry teaching and learning: Philosophical considerations. In *International handbook of research in history, philosophy and science teaching* (pp. 1363–1380). Springer.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., and Keselman, J. C. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research*, *68*(3), 350–386.

Khoiriyah, A. J., and Husamah, H. (2018). Problem-based learning: Creative thinking skills, problem-solving skills, and learning outcome of seventh grade students. *JPBI (Jurnal Pendidikan Biologi Indonesia)*, *4*(2), 151–160.

Kiers, H. A., and Mechelen, I. V. (2001). Three-way component analysis: Principles and illustrative application. *Psychological Methods*, *6*(1), 84.

Kim, N. J., Belland, B. R., and Walker, A. E. (2018). Effectiveness of computer-based scaffolding in the context of problem-based learning for STEM education: Bayesian meta-analysis. *Educational Psychology Review*, *30*(2), 397–429.

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why unguided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. *Educational Psychologist*, *41*(2), 75–86.

Kirschner, P. A., Sweller, J., Kirschner, F., and Zambrano R, J. (2018). From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning*, *13*(2), 213–233.

Klahr, D., and Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science*, *15*(10), 661–667.

Kyun, S., Kalyuga, S., and 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.

Lau, K., and Lam, T. Y. (2017). Instructional practices and science performance of 10 top-performing regions in PISA 2015. *International Journal of Science Education*, *39*(15), 2128–2149.

Lavonen, J., and Laaksonen, S. (2009). Context of teaching and learning school science in Finland: Reflections on PISA 2006 results. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *46*(8), 922–944.

Leary, H., Walker, A., Shelton, B. E., and Fitt, M. H. (2013). Exploring the relationships between tutor background, tutor training, and student learning: A problem-based learning meta-analysis. *Interdisciplinary Journal of Problem-Based Learning*, *7*(1), 40–66.

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., and Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, *45*, 1058–1072.

Leppink, J., Paas, F., Van Gog, T., Der Vleuten, C. P. van, and Van Merrienboer, J. J. (2014). Effects of pairs of problems and examples on task performance and different types of cognitive load. *Learning and Instruction*, *30*, 32–42.

Lesaffre, E., Rizopoulos, D., and Tsonaka, R. (2007). The logistic transform for bounded outcome scores. *Biostatistics*, *8*(1), 72–85.

Lespiau, F., and Tricot, A. (2018). Primary knowledge enhances performance and motivation in reasoning. *Learning and Instruction*, *56*, 10–19.

Lespiau, F., and Tricot, A. (2019). Using primary knowledge: An efficient way to motivate students and promote the learning of formal reasoning. *Educational Psychology Review*, *31*(4), 915–938.

Li, B., and Walder, A. G. (2001). Career advancement as party patronage: Sponsored mobility into the Chinese administrative elite, 1949–1996. *American Journal of Sociology*, *106*(5), 1371–1408.

Likourezos, V., and Kalyuga, S. (2017). Instruction-first and problem-solving-first approaches: Alternative pathways to learning complex tasks. *Instructional Science*, *45*(2), 195–219.

Liou, P.-Y. (2021). Students' attitudes toward science and science achievement: An analysis of the differential effects of science instructional practices. *Journal of Research in Science Teaching*, *58*(3), 310–334.

Liou, P.-Y., and Jessie Ho, H.-N. (2018). Relationships among instructional practices, students' motivational beliefs and science achievement in Taiwan using hierarchical linear modelling. *Research Papers in Education*, *33*(1), 73–88.

Lipsey, M. W., and Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, *48*(12), 1181.

Liu, Y. (2013). Meritocracy and the Gaokao: A survey study of higher education selection and socio-economic participation in East China. *British Journal of Sociology of Education*, *34*(5-6), 868–887.

Liu, Y. (2016). *Higher education, meritocracy and inequality in China*. Springer.

Loibl, K., Roll, I., and Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*(4), 693–715.

Loibl, K., Tillema, M., Rummel, N., and Gog, T. van. (2020). The effect of contrasting cases during problem solving prior to and after instruction. *Instructional Science*, *48*(2), 115–136.

Loyens, S. M., Jones, S. H., Mikkers, J., and Gog, T. van. (2015). Problem-based learning as a facilitator of conceptual change. *Learning and Instruction*, *38*, 34–42.

Lu, J., Kalyuga, S., and Sweller, J. (2020). Altering element interactivity and variability in example-practice sequences to enhance learning to write Chinese characters. *Applied Cognitive Psychology*, *34*(4), 837–843.

Martin, A. J. (2016). *Using Load Reduction Instruction (LRI) to boost motivation and engagement*. British Psychological Society Leicester.

Martin, A. J., and Evans, P. (2018). Load reduction instruction: Exploring a framework that assesses explicit instruction through to independent learning. *Teaching and Teacher Education*, *73*, 203–214.

Martin, A. J., and Evans, P. (2019). Load reduction instruction (LRI): 14Sequencing explicit instruction and guided discovery to enhance students' motivation, engagement, learning, and achievement. In *Advances in cognitive load theory* (pp. 15–29). Routledge.

Matlen, B. J., and Klahr, D. (2013). Sequential effects of high and low instructional guidance on children's acquisition of experimentation skills: Is it all in the timing? *Instructional Science*, *41*(3), 621–634.

Matthews, M. R. (2015). Reflections on 25 years of journal editorship. *Science & Education*, *24*(5), 749–805.

McConney, A., Oliver, M. C., Woods-Mcconney, A., Schibeci, R., and Maor, D. (2014). Inquiry, engagement, and literacy in science: A retrospective, cross-national analysis using PISA 2006. *Science Education*, *98*(6), 963–980.

McDonald, R. P. (2013). *Test theory: A unified treatment*. psychology press.

Merritt, J., Lee, M. Y., Rillero, P., and Kinach, B. M. (2017). Problem-based learning in K–8 mathematics and science education: A literature review. *Interdisciplinary Journal of Problem-Based Learning*, *11*(2).

Minner, D. D., Levy, A. J., and Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *47*(4), 474–496.

Moallem, M. (2019). Effects of PBL on Learning Outcomes, Knowledge Acquisition, and Higher-Order Thinking Skills. *The Wiley Handbook of Problem-Based Learning*, 107–133.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, *32*(1), 99–113.

Mostafa, T., Echazarra, A., and Guillou, H. (2018). *The science of teaching science: An exploration of science teaching practices in PISA 2015*.

Muthanna, A., and Sang, G. (2016). Undergraduate Chinese students' perspectives on Gaokao examination: Strengths, weaknesses, and implications. *International Journal of Research Studies in Education*, *5*(2), 3–12.

Nachtigall, V., Serova, K., and Rummel, N. (2020). When failure fails to be productive: Probing the effectiveness of productive failure for learning beyond STEM domains. *Instructional Science*, *48*(6), 651–697.

Nanda, B., and Manjunatha, S. (2013). Indian medical students' perspectives on problem-based learning experiences in the undergraduate curriculum: One size does not fit all. *J Educ Eval Health Prof*, *10*(1), 11.

Neville, A., Norman, G., and White, R. (2019). McMaster at 50: Lessons learned from five decades of PBL. *Advances in Health Sciences Education*, *24*(5), 853–863.

Newman, P., and DeCaro, M. S. (2018). How Much Support Is Optimal During Exploratory Learning? *CogSci*.

Novak, A. M., and Krajcik, J. S. (2019). A case study of project-based learning of middle school students exploring water quality. *The Wiley Handbook of Problem-Based Learning*, 551–572.

Nyumba, T. O., Wilson, K., Derrick, C. J., and Mukherjee, N. (2018). The use of focus group discussion methodology: Insights from two decades of application in conservation. *Methods in Ecology and Evolution*, *9*(1), 20–32.

O'Brien, R. G., and Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, *97*(2), 316.

OECD. (2011). *Lessons from PISA for the United States: Strong performers and successful reformers in education*. OECD Publishing. https://doi.org/10.1787/9789264096660-en

OECD. (2009). *PISA 2006 Technical Report*. https://www.oecd-ilibrary.org/content/publication/9789264048096-en

OECD. (2017). *PISA 2015 Technical Background*. https://www.oecd-ilibrary.org/content/component/9789264285521-14-en

Oliver, M., McConney, A., and Woods-McConney, A. (2021). The efficacy of inquiry-based instruction in science: A comparative analysis of six countries using PISA 2015. *Research in Science Education*, *51*(2), 595–616.

Onyon, C. (2012). Problem-based learning: A review of the educational and psychological theory. *The Clinical Teacher*, *9*(1), 22–26.

Ouwehand, K., Kroef, A. van der, Wong, J., and Paas, F. (2022). Measuring cognitive load: Are there more valid alternatives to likert rating scales? *Frontiers in Education*, *6*, 702616.

Paine, L. (1992). Teaching and modernization in contemporary China. *Education and Modernization: The Chinese Experience*, 183–209.

Park, B., Moreno, R., Seufert, T., and Brünken, R. (2011). Does cognitive load moderate the seductive details effect? A multimedia study. *Computers in Human Behavior*, *27*(1), 5–10.

Pavlik, J. V. (2023). Collaborating With ChatGPT: Considering the Implications of Generative Artificial Intelligence for Journalism and Media Education. *Journalism & Mass Communication Educator*, 10776958221149577.

Pekrun, R., and Linnenbrink-Garcia, L. (2012). Academic emotions and student engagement. In *Handbook of research on student engagement* (pp. 259–282). Springer.

Penjvini, S., and Shahsawari, S. S. (2013). Comparing problem based learning with lecture based learning on medicine giving skill to newborn in nursing students. *Journal of Nursing Education and Practice*, *3*(9), 53.

Plass, J. L., and Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review*, *31*(2), 339–359.

Rehmat, A. P., and Hartley, K. (2020). Building engineering awareness: Problem based learning approach for STEM integration. *Interdisciplinary Journal of Problem-Based Learning*, *14*(1).

Renken, M. D., and Nunez, N. (2010). Evidence for improved conclusion accuracy after reading about rather than conducting a belief-inconsistent simple physics experiment. *Applied Cognitive Psychology*, *24*(6), 792–811.

Revelle, W., and Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154.

Rey, G. D., and Steib, N. (2013). The personalization effect in multimedia learning: The influence of dialect. *Computers in Human Behavior*, *29*(5), 2022–2028.

Richey, J. E., and Nokes-Malach, T. J. (2013). How much is too much? Learning and motivation effects of adding instructional explanations to worked examples. *Learning and Instruction*, *25*, 104–124.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*(1), 1–15.

Robinson, L., Harris, A., and Burton, R. (2015). Saving face: Managing rapport in a problem-based learning group. *Active Learning in Higher Education*, *16*(1), 11–24.

Roll, I., Butler, D., Yee, N., Welsh, A., Perez, S., Briseno, A., Perkins, K., and Bonn, D. (2018). Understanding the impact of guiding inquiry: The relationship between directive support, student attributes, and transfer of knowledge, attitudes, and behaviours in inquiry learning. *Instructional Science*, *46*(1), 77–104.

Russell, B., and Linsky, B. (2020). *The problem of China*. Routledge.

Russo, J., and Hopkins, S. (2019). Teaching primary mathematics with challenging tasks: How should lessons be structured? *The Journal of Educational Research*, *112*(1), 98–109.

Ryan, J., Kang, C., Mitchell, I., and Erickson, G. (2009). China's basic education reform: An account of an international collaborative research and development project. *Asia Pacific Journal of Education*, *29*(4), 427–441.

Sargent, T. C. (2015). Professional learning communities and the diffusion of pedagogical innovation in the Chinese education system. *Comparative Education Review*, *59*(1), 102–132.

Savery, J. R. (2015). Overview of problem-based learning: Definitions and distinctions. *Essential Readings in Problem-Based Learning: Exploring and Extending the Legacy of Howard S. Barrows*, *9*(2), 5–15.

Savery, J. R. (2019). Comparative pedagogical models of problem-based learning. *The Wiley Handbook of Problem-Based Learning*, 81–104.

Scheiter, K., Schubert, C., Gerjets, P., and Stalbovs, K. (2015). Does a strategy training foster students' ability to learn from multimedia? *The Journal of Experimental Education*, *83*(2), 266–289.

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., and Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science*, *43*(1), 93–114.

Schmidt, H. G. (2012). A brief history of problem-based learning. In *One-day, one-problem* (pp. 21–40). Springer.

Schmidt, H. G., Loyens, S. M., Van Gog, T., and Paas, F. (2007). Problem-based learning is compatible with human cognitive architecture: Commentary on Kirschner, Sweller, and. *Educational Psychologist*, *42*(2), 91–97.

Schmidt, H. G., and Mamede, S. (2020). How cognitive psychology changed the face of medical education research. *Advances in Health Sciences Education*, *25*(5), 1025–1043.

Schmidt, H. G., Rotgans, J. I., and Yew, E. H. J. (2019). Cognitive Constructivist Foundations of Problem-Based Learning. *The Wiley Handbook of Problem-Based Learning*, 25–50.

Schmidt, H. G., Van der Molen, H. T., Te Winkel, W. W., and Wijnen, W. H. (2009). Constructivist, problem-based learning does work: A meta-analysis of curricular comparisons involving a single medical school. *Educational Psychologist*, *44*(4), 227–249.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., and Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759.

Schweisfurth, M. (2013). *Learner-centred education in international perspective: Whose pedagogy for whose development?* Routledge.

Seibert, S. A. (2021). Problem-based learning: A strategy to foster generation Z's critical thinking and perseverance. *Teaching and Learning in Nursing*, *16*(1), 85–88.

Sepp, S., Howard, S. J., Tindall-Ford, S., Agostinho, S., and Paas, F. (2019). Cognitive load theory and human movement: Towards an integrated model of working memory. *Educational Psychology Review*, *31*(2), 293–317.

Servant-Miklos, V. F. (2019a). A Revolution in its own right: How maastricht university reinvented problem-based learning. *Health Professions Education*, *5*(4), 283–293.

Servant-Miklos, V. F. (2019b). Fifty years on: A retrospective on the world's first problem-based learning programme at McMaster University Medical School. *Health Professions Education*, *5*(1), 3–12.

Servant-Miklos, V. F. (2019c). Problem solving skills versus knowledge acquisition: The historical dispute that split problem-based learning into two camps. *Advances in Health Sciences Education*, *24*(3), 619–635.

Servant-Miklos, V. F., Norman, G. R., and Schmidt, H. G. (2019). A short intellectual history of problem-based learning. *The Wiley Handbook of Problem-Based Learning*, 3–24.

Servant-Miklos, V. F., and Spliid, C. M. (2017). The construction of teaching roles at Aalborg university centre, 1970–1980. *History of Education*, *46*(6), 788–809.

Servant-Miklos, V. F., Woods, N. N., and Dolmans, D. H. (2019). *Celebrating 50 years of problem-based learning: Progress, pitfalls and possibilities*. Springer.

Simamora, R. E., Sidabutar, D. R., and Surya, E. (2017). Improving learning activity and students' problem solving skill through problem based learning (PBL) in junior high school. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, *33*(2), 321–331.

Sinha, T., and Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. *Review of Educational Research*, *91*(5), 761–798.

Slavin, R. E. (2014). Cooperative Learning and Academic Achievement: Why Does Groupwork Work?.[Aprendizaje cooperativo y rendimiento académico:?' Por qué funciona el trabajo en grupo?]. *Anales de Psicología/Annals of Psychology*, *30*(3), 785–791.

Stage, F. K., Carter, H. C., and Nora, A. (2004). Path analysis: An introduction and analysis of a decade of research. *The Journal of Educational Research*, *98*(1), 5–13.

Stasavage, D. (2020). The decline and rise of democracy. In *The Decline and Rise of Democracy*. Princeton University Press.

Steenhof, N., Woods, N. N., and Mylopoulos, M. (2020). Exploring why we learn from productive failure: Insights from the cognitive and learning sciences. *Advances in Health Sciences Education*, *25*(5), 1099–1106.

Stull, A. T., and Mayer, R. E. (2007). Learning by doing versus learning by viewing: Three experimental comparisons of learner-generated versus author-provided graphic organizers. *Journal of Educational Psychology*, *99*(4), 808.

Suh, J. M., and Seshaiyer, P. (2019). Promoting Ambitious Teaching and Learning through Implementing Mathematical Modeling in a PBL Environment: A Case Study. *The Wiley Handbook of Problem-Based Learning*, 529–550.

Sweller, J. (2009). What human cognitive architecture tells us about constructivism. In *Constructivist Instruction* (pp. 139–155). Routledge.

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16.

Sweller, J. (2021). The role of evolutionary psychology in our understanding of human cognition: Consequences for cognitive load theory and instructional procedures. *Educational Psychology Review*, 1–13.

Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory*. Springer.

Sweller, J., Kirschner, P. A., and Clark, R. E. (2007). Why minimally guided teaching techniques do not work: A reply to commentaries. *Educational Psychologist*, *42*(2), 115–121.

Sweller, J., Merriënboer, J. J. van, and Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292.

Taber, K. S. (2013). *Modeling learners and learning in science education*. Springer.

Tan, C. (2012). *Learning from Shanghai: Lessons on achieving educational success* (Vol. 21). Springer Science & Business Media.

Tan, C. (2016). *Educational policy borrowing in China: Looking West or looking East?* Routledge.

Tawfik, A. A., Hung, W., and Giabbanelli, P. J. (2020). Comparing How Different Inquiry-Based Approaches Impact Learning Outcomes. *Interdisciplinary Journal of Problem-Based Learning*, *14*(1), n1.

Tawfik, A. A., Rong, H., and Choi, I. (2015). Failing to learn: Towards a unified design approach for failure-based learning. *Educational Technology Research and Development*, *63*(6), 975–994.

Teig, N., Scherer, R., and Nilsen, T. (2018). More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science. *Learning and Instruction*, *56*, 20–29.

Thoemmes, F. J., and Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research*, *46*(1), 90–118.

Tobias, S., and Duffy, T. M. (2009). The success or failure of constructivist instruction: An introduction. In *Constructivist Instruction* (pp. 15–22). Routledge.

Tobias, S., Kirschner, P. A., Rosenshine, B. V., Jonassen, D. H., and Spiro, R. J. (2007). Debate: Constructivism, discovery, problem-based, experiential, and inquiry-based teaching–Success or failure. *American Educational Research Association*.

UNESCO, U. (2019). Global flow of tertiary-level students. *UNESCO Institute for Statistics*.

Van Merrienboer, J. J., and Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177.

Vickers, E., and Xiaodong, Z. (2017). *Education and society in post-Mao China*. Routledge.

Wang, J., and Buck, G. (2015). The relationship between Chinese students' subject matter knowledge and argumentation pedagogy. *International Journal of Science Education*, *37*(2), 340–366.

Weaver, J. P., Chastain, R. J., DeCaro, D. A., and DeCaro, M. S. (2018). Reverse the routine: Problem solving before instruction improves conceptual knowledge in undergraduate physics. *Contemporary Educational Psychology*, *52*, 36–47.

Wei, S.-J., Xie, Z., and Zhang, X. (2017). From" Made in China" to" Innovated in China": Necessity, prospect, and challenges. *Journal of Economic Perspectives*, *31*(1), 49–70.

Westhues, A., Barsen, C., Freymond, N., and Train, P. (2014). An outcome evaluation of a problem-based learning approach with MSW students. *Journal of Social Work Education*, *50*(3), 472–489.

Wijnen, M., Loyens, S. M., Wijnia, L., Smeets, G., Kroeze, M. J., and Van der Molen, H. T. (2018). Is problem-based learning associated with students' motivation? A quantitative and qualitative study. *Learning Environments Research*, *21*(2), 173–193.

Wijnia, L., Loyens, S. M., Noordzij, G., Arends, L. R., and Rikers, R. (2017). The effects of problem-based, project-based, and case-based learning on students' motivation: A meta-analysis. *Eindrapport NRO-Project*, 405–415.

Wijnia, L., Loyens, S. M., and Rikers, R. M. (2019). The problem-based learning process: An overview of different models. *The Wiley Handbook of Problem-Based Learning*, 273–295.

Wise, A. F., and O'Neill, K. (2009). Beyond more versus less: A reframing of the debate on instructional guidance. In *Constructivist instruction* (pp. 94–117). Routledge.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Wright, S. (1960). The treatment of reciprocal interaction, with or without lag, in path analysis. *Biometrics*, *16*(3), 423–445.

Xu, K. M., Koorn, P., De Koning, B., Skuballa, I. T., Lin, L., Henderikx, M., Marsh, H. W., Sweller, J., and Paas, F. (2021). A growth mindset lowers perceived cognitive load and improves learning: Integrating motivation to cognitive load. *Journal of Educational Psychology*, *113*(6), 1177.

Xue, M. M. (2021). Autocratic rule and social capital: Evidence from imperial China. *Available at SSRN 2856803*.

Yew, E. H., and Goh, K. (2016). Problem-based learning: An overview of its process and impact on learning. *Health Professions Education*, *2*(2), 75–79.

You, Y. (2019). The seeming 'round trip'of learner-centred education: A 'best practice'derived from China's New Curriculum Reform? *Comparative Education*, *55*(1), 97–115.

Zambrano, J., Kirschner, F., Sweller, J., and Kirschner, P. A. (2019a). Effects of group experience and information distribution on collaborative learning. *Instructional Science*, *47*(5), 531–550.

Zambrano, J., Kirschner, F., Sweller, J., and Kirschner, P. A. (2019b). Effects of prior knowledge on collaborative and individual learning. *Learning and Instruction*, *63*, 101214.

Zhai, X. (2022). ChatGPT user experience: Implications for education. *Available at SSRN 4312418*.

Zhang, L. (2016). Is inquiry-based science teaching worth the effort? *Science & Education*, *25*(7), 897–915.

Zhang, L. (2018). Withholding answers during hands-on scientific investigations? Comparing effects on developing students' scientific knowledge, reasoning, and application. *International Journal of Science Education, 40*(4), 459–469.

Zhang, L. (2019). "Hands-on" plus "inquiry?" Effects of withholding answers coupled with physical manipulations on students' learning of energy-related science concepts. *Learning and Instruction, 60,* 199–205.

Zhang, L., and Cobern, W. W. (2021). Confusions on "guidance" in inquiry-based science teaching: A response to Aditomo and Klieme (2020). *Canadian Journal of Science, Mathematics and Technology Education, 21*(1), 207–212.

Zhang, L., Kirschner, P. A., Cobern, W. W., and Sweller, J. (2022). There is an evidence crisis in science educational policy. *Educational Psychology Review, 34*(2), 1157–1176.

Zhang, L., and Li, Z. (2019). How does inquiry-based scientific investigation relate to the development of students' science knowledge, knowing, applying, and reasoning? An examination of TIMSS data. *Canadian Journal of Science, Mathematics and Technology Education, 19*(3), 334–345.

Zhang, W., and Bray, M. (2017). Micro-neoliberalism in China: Public-private interactions at the confluence of mainstream and shadow education. *Journal of Education Policy, 32*(1), 63–81.

Zlomuzica, A., Preusser, F., Totzeck, C., Dere, E., and Margraf, J. (2016). The impact of different emotional states on the memory for what, where and when features of specific events. *Behavioural Brain Research, 298,* 181–187.

Appendices

Appendix 1 Acronym glossary of the present thesis

Acronym	Full form	Location
ANCOVA	Analysis of Covariance	Chapter 4
ANOVA	Analysis of Variance	Chapter 4
CCLT	Collaborative cognitive load theory	Section 2.2.4
CFA	Confirmatory factor analysis	Section 2.3.1.1
CLT	Cognitive load theory	Chapter 1
ICLT	The interval theory view of cognitive load theory	Section 2.2.4
LHZ School	Luzhou Huojing Zhan School	Chapter 4
LMM	linear mixed-effects model	Chapter 4
MANCOVA	Multivariate Analysis of Covariance	Chapter 4
MANOVA	Multivariate Analysis of Variance	Chapter 4
ML	Maximum likelihood	Chapter 4
PBL	Problem-based learning	Chapter 1
PISA	Programme for International Student Assessment	Section 2.3.1.1
REML	Restricted maximum likelihood	Chapter 4
RZS School	Ruian Zijing Shuyuan Schoo	Chapter 4
SEM	Structural Equation Modeling	Section 3.1.2
STEM	The disciplines of science, technology, engineering, and mathematics	Section 3.1.1
TIMSS	Trends in International Mathematics and Science Study	Section 2.3.1.1

Source: Compiled by author

Appendix 2 Technical details of standardizing previous empirical findings on PBL efficacy

The original coefficients and standard errors from studies are not comparable. To put various experiments or analyses onto a level playing field, the present thesis standardizes those statistics in the following steps.

1) Convert all the coefficient and standard errors of the treatment (guidance, early guidance, and group discussion respectively) into incremental form. If the original statistics are already in incremental form, there is no change for them in this step. For the original statistics in the raw-value form, the present thesis calculates the incremental effect and standard error using the formulas below.

$$Incremental\ Effect = Mean_{Treatment} - Mean_{Control}$$

$$SE\ of\ Incremental\ Effect = \sqrt{\frac{(N_{Treatment} - 1)SE_{Treatment}^2 + (N_{Control} - 1)SE_{Control}^2}{N_{Treatment} + N_{Contrl} - 2}}$$

The right side of the equation calculating SE of incremental effect is the denominator of Cohen's d. The studies reviewed by the present thesis apply the condition of using Cohen's d since 1) there are no significant differences in sample sizes between treatment and control groups and 2) their sample sizes are greater than 20 (J. Cohen, 2013).

2) Scale all the incremental effect coefficients and standard errors so their standard errors are of identical value.

The limitations of the standardization procedure used here include that it will understate the incremental effect if the study inherently has a larger SE. In terms of testing the hypothesis, both larger SE and small estimated effect impede rejecting the null hypothesis. However, it is worth noting that the standardized value of incremental effect in Section 2.3 of the present thesis should not be inferred as the magnitude level of the incremental effect of treatment. Instead, it is the blend of effect and standard error.

Appendix 3 Definition of PISA items

PISA Version	PISA Item Code	Item Level	Explanation
PISA 2006	SCINTACT	1	Item parameters for science teaching: interaction
	ST34Q01	2	a) Students are given opportunities to explain their ideas
	ST34Q05	2	e) The lessons involve students' opinions about the topics
	ST34Q09	2	i) There is a class debate or discussion
	ST34Q13	2	m) The students have discussions about the topics
	SCHANDS	1	Item parameters for science teaching: hands-on activities
	ST34Q02	2	b) Students spend time in the laboratory doing practical experiments
	ST34Q03	2	c) Students are required to design how a question could be investigated in the laboratory
	ST34Q06	2	f) Students are asked to draw conclusions from an experiment they have conducted
	ST34Q14	2	n) Students do experiments by following the instructions of the teacher
	SCINVEST	1	Item parameters for science teaching: student investigations
	ST34Q08	2	h) Students are allowed to design their own experiments
	ST34Q11	2	k) Students are given the chance to choose their own investigations
	ST34Q16	2	p) Students are asked to do an investigation to test out their own ideas
	SCAPPLY	1	Item parameters for science teaching: focus on models or applications
	ST34Q07	2	g) The teacher explains how a idea can be applied to a number of different phenomena (e.g. the movement of objects, substances with similar properties)
	ST34Q12	2	l) The teacher uses science to help students understand the world outside school
	ST34Q15	2	o) The teacher clearly explains the relevance of concepts to our lives
	ST34Q17	2	q) The teacher uses examples of technological application to show how is relevant to society
PISA 2015	IBTEACH	1	Inquiry-based science teaching an learning practices (WLE)
	ST098Q01TA	2	Students are given opportunities to explain their ideas.
	ST098Q02TA	2	Students spend time in the laboratory doing practical experiments.
	ST098Q03NA	2	Students are required to argue about science questions.
	ST098Q05TA	2	Students are asked to draw conclusions from an experiment they have conducted.
	ST098Q06TA	2	The teacher explains <school science> idea can be applied

PISA Version	PISA Item Code	Item Level	Explanation
	ST098Q07 TA	2	Students are allowed to design their own experiments.
	ST098Q08 NA	2	There is a class debate about investigations.
	ST098Q09 TA	2	The teacher clearly explains relevance <broad science> concepts to our lives.
	ST098Q10 NA	2	Students are asked to do an investigation to test ideas.
	TDTEAC H	1	Teacher-directed science instruction (WLE)
	ST103Q01 NA	2	The teacher explains scientific ideas.
	ST103Q03 NA	2	A whole class discussion takes place with the teacher.
	ST103Q08 NA	2	The teacher discusses our questions.
	ST103Q11 NA	2	The teacher demonstrates an idea.

Source: OECD (2009) and OECD (2017), compiled by author

Appendix 4 Experimental studies of PBL by education disciplines

Education discipline	Studies	# of studies	% of studies
Science	Hsu et al. (2015), Klahr and Nigam (2004), Stull and Mayer (2007), Zhang (2019), Zhang (2018), Ashman et al. (2020), Hsu et al. (2015), Chase and Klahr (2017), Matlen and Klahr (2013), Chen et al. (2021), Nachtigall et al. (2020), Weaver et al. (2018), Aidoo et al. (2016), Argaw et al. (2016)	14	37.8
Mathematics	Rittle-Johnson (2006), Chen et al. (2016), Chen et al. (2019), Chen et al. (2020), Likourezos and Kalyuga (2017), Loibl et al. (2020), Kapur (2014), Kazemi and Ghoraishi (2012), Ajai et al. (2013), Firdaus and Herman (2017), Hendriana et al. (2018), Amalia et al. (2017)	12	32.4
Medical	Steenhof et al. (2020), Imanieh et al. (2014), Penjvini and Shahsawari (2013)	3	8.1
Education	Barth et al. (2019), Hendarwati et al. (2021)	2	5.4
Language	Kyun et al. (2013), Lu et al. (2020)	2	5.4
Engineering	Jabarullah and Hussain (2019)	1	2.7
Not identified	De Witte and Rogge (2016)	1	2.7
Psychology	Moreno (2004)	1	2.7
Social work	Westhues et al. (2014)	1	2.7

Source: Compiled by author

Appendix 5 Regression results of the expected learning outcomes

In order to calculate the anticipated outcomes of learning for the stratified random assignment, the present study employs a linear mixed model (LMM) as outlined in Section 4.1. The parameters of Section 4.1 will be estimated using a restricted maximum likelihood (REML) technique, as opposed to the commonly used standard maximum likelihood (ML) method. The REML approach is deemed more suitable for small-sample analysis (Galecki et al. (2013) provide a comprehensive understanding of linear mixed models and REML). The results of the LMM model for each experiment are presented below.

	Experiment 1	**Experiment 2**	**Experiment 3**	**Experiment 4**	**Experiment 5**	**Experiment 6**
Intercept	57.354*** (12.01)	58.712*** (10.58)	53.228*** (13.69)	69.668*** (11.97)	38.644*** (10.75)	81.701*** (9.16)
Gender	-1.541*** (0.51)	-0.560 (0.46)	-1.007* (0.61)	-0.700 (0.53)	-0.641 (0.47)	-0.749* (0.42)
Age	1.168 (0.84)	1.318* (0.76)	1.602 (1.00)	1.195 (0.89)	2.533*** (0.77)	0.196 (0.68)
SESI	-0.004 (0.31)	-0.270 (0.30)	0.802* (0.42)	-0.256 (0.35)	0.400 (0.31)	0.205 (0.27)
Random effect of Class	Controlled	Controlled	Controlled	Controlled	Controlled	Controlled
N	528	477	209	223	453	236

Source: Compiled by author

The above table presents the LMM results for all six experiments of this thesis. The model includes three predictors: $Gender$, Age, and $SESI$, as well as a random effect of $Class$, which is controlled in all six experiments. The coefficients for each predictor are reported along with their standard errors in parentheses and significance levels indicated by asterisks (* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$). The number of observations for each experiment is also reported.

Appendix 6 Socio-economic status questions

Item SES1: What is your father's occupation?

- Managerial or professional
- Technical, associate professional or administrative
- Skilled manual worker
- Semiskilled or unskilled manual worker
- Not working or other

Item SES2: What is your mother's occupation?

- Managerial or professional
- Technical, associate professional or administrative
- Skilled manual worker
- Semiskilled or unskilled manual worker
- Not working or other

Item SES3: What is your parents' highest level of education?

- Both have completed tertiary education
- One has completed tertiary education and the other has completed upper secondary education
- Both have completed upper secondary education
- One has completed upper secondary education and the other has completed lower secondary education
- Both have completed lower secondary education or less

Item SES4: How many books are there in your home?

- None
- 1-10
- 11-25
- 26-50

- 51 or more

Item SES5: How often do you eat dinner together with your family?

- Almost every day
- Once or twice a week
- A few times a month
- Almost never

Item SES6: How often do you take part in cultural activities (e.g. going to the cinema, theatre or concert)?

- Once a week or more
- Once or twice a month
- A few times a year
- Almost never

Item SES7: How many computers are there at home for your own use?

- None
- 1
- 2
- 3 or more

Item SES8: How many rooms are there in your home for sleeping?

- 1
- 2
- 3
- 4 or more

Item SES9: How many people sleep in your room?

- 1
- 2
- 3

- 4 or more

Item SES10: How often do you have to go without eating because there is not enough food at home?

- Never
- Rarely
- Sometimes
- Often

To transfer the answers into a numeric measure, a scoring system can be used. This system assigns a specific numerical value to each answer option, allowing for the data to be analyzed quantitatively. For example, for the question about the number of books in the home, a scoring system might assign the following values:

- None = 1
- 1-10 = 2
- 11-25 = 3
- 26-50 = 4
- 51 or more = 5

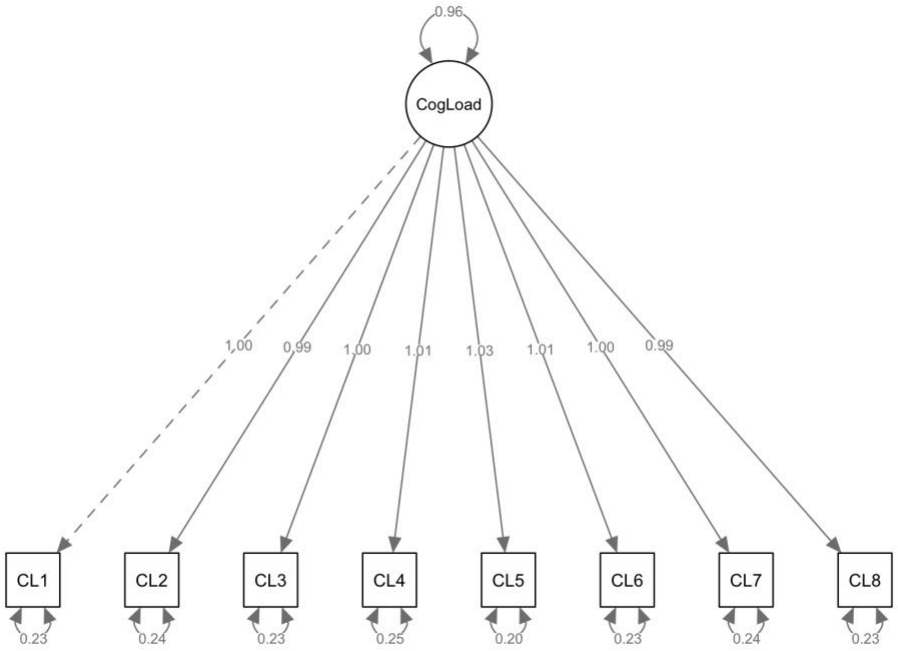
The final SESI is calculated by the following equation.

$$SESI = \sum_{i=1}^{4} SES_i/5 + \sum_{j=5}^{10} SES_j/4$$

Appendix 7 Constructing CFA latent variables

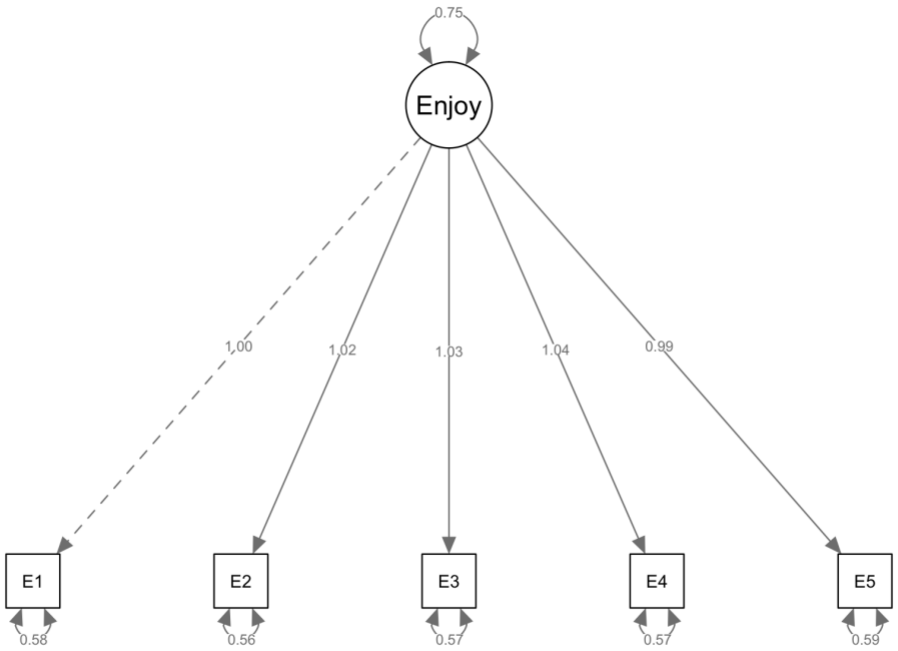
Figures Appendix 7.1, Appendix 7.2, and Appendix 7.3 present the CFA latent variable models for Cognitive Load, Enjoyment, and Self-efficacy, respectively. The factor loading of the first item is standardized to one, thereby expressing the magnitude of the other factor loadings relative to it, facilitating the interpretation of the size and significance of the other loadings. Furthermore, this normalization helps to establish a standardized metric for the latent variable, enabling comparisons with other latent variables to be made with greater ease.

Figure Appendix 7.1



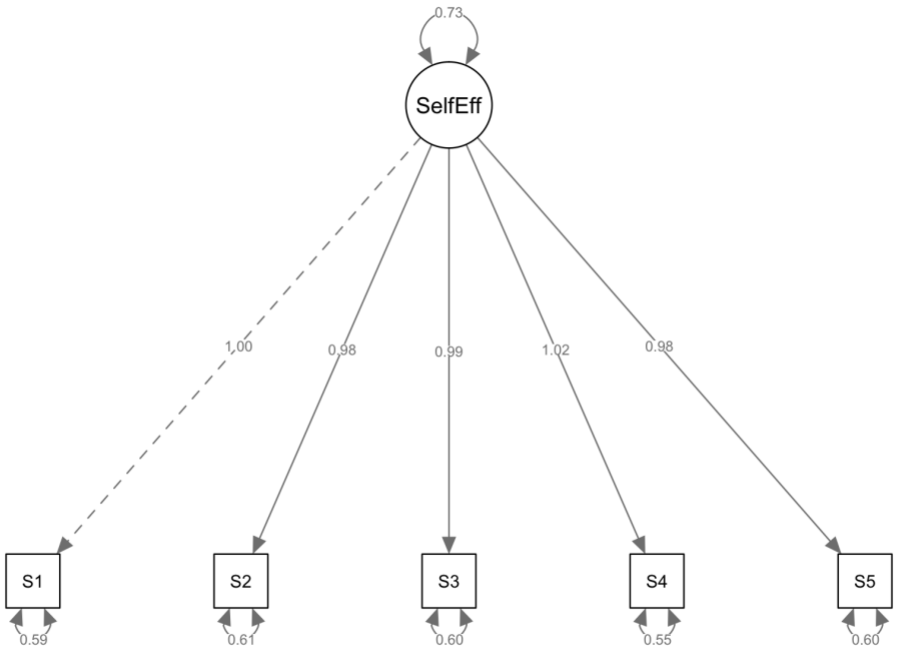
Source: Compiled by author

Figure Appendix 7.2



Source: Compiled by author

Figure Appendix 7.3



Source: Compiled by author

Appendix 8 Cognitive load questions

Item CL1: How mentally demanding did you find the task?

- Very demanding
- Somewhat demanding
- Not very demanding
- Not at all demanding

Item CL2: How much effort did you have to put into the task?

- A lot of effort
- Some effort
- A little effort
- No effort

Item CL3: How much did you have to focus on the task?

- All the time
- Most of the time
- Some of the time
- Not at all

Item CL4: How difficult did you find the task?

- Very difficult
- Somewhat difficult
- Not very difficult
- Not at all difficult

Item CL5: How much did you feel you were working at your limit?

- All the time
- Most of the time
- Some of the time

- Not at all

Item CL6: How much did you feel you were using your cognitive resources?

- All the time
- Most of the time
- Some of the time
- Not at all

Item CL7: How much did you feel the task was taking your attention away from other things?

- All the time
- Most of the time
- Some of the time
- Not at all

Item CL8: How much did you feel that the task was interfering with other things you were doing?

- All the time
- Most of the time
- Some of the time
- Not at all

Appendix 9 Enjoyment questions

How much do you agree with the statements below?

Item E1: I generally have fun when I am learning Physics topics.

- Strongly disagree
- Disagree
- Agree
- Strongly agree

Item E2: I like reading about Physics.

- Strongly disagree
- Disagree
- Agree
- Strongly agree

Item E3: I am happy doing Physics problems.

- Strongly disagree
- Disagree
- Agree
- Strongly agree

Item E4: I enjoy acquiring new knowledge in Physics.

- Strongly disagree
- Disagree
- Agree
- Strongly agree

Item E5: I am interested in learning about Physics.

- Strongly disagree
- Disagree

- Agree
- Strongly agree

Appendix 10 Self-efficacy questions

Item S1: How confident are you in your ability to understand and apply Newton's laws of motion in a physics problem?

- Very confident
- Somewhat confident
- Not very confident
- Not at all confident

Item S2: How confident are you in your ability to solve problems related to Newton's laws of motion?

- Very confident
- Somewhat confident
- Not very confident
- Not at all confident

Item S3: How confident are you in your ability to explain Newton's laws of motion to others?

- Very confident
- Somewhat confident
- Not very confident
- Not at all confident

Item S4: How confident are you in your ability to use Newton's laws of motion to predict the behavior of objects in motion?

- Very confident
- Somewhat confident
- Not very confident
- Not at all confident

Item S5: How confident are you in your ability to apply Newton's laws of motion to real-world situations?

- Very confident
- Somewhat confident
- Not very confident
- Not at all confident

Appendix 11 Maximum guidance description

Opening: “Hello and welcome to this video lecture on Newton’s laws. My name is [Name] and I’ll be guiding you through the three laws that explain the motion of objects. By the end of this video, you will have a solid understanding of how these laws work and how they apply to the world around us. Let’s get started.”

Law 1: “The first law is known as the law of inertia. It states that an object at rest will remain at rest, and an object in motion will remain in motion at a constant velocity unless acted upon by an unbalanced force. Let’s take a look at an example. Imagine a hockey puck on a frozen lake. The hockey puck is at rest and there are no unbalanced forces acting upon it. According to the law of inertia, the hockey puck will remain at rest. Now, imagine a player hits the hockey puck with a stick. This action is an unbalanced force, and the hockey puck will start to move. This is how the law of inertia works.

Let us proceed by presenting a step-by-step worked example for further illustration.” (Insert one step-by-step worked example)

Law 2: “The second law is known as the force-mass-acceleration relationship. It states that force is equal to mass multiplied by acceleration. In other words, F = ma. This means that if you want to accelerate an object, you need to apply a force. The greater the mass of the object, the more force is needed to accelerate it. Let’s take a look at an example. Imagine you’re pushing a shopping cart. The cart has a mass of 20 kg and you’re applying a force of 10 N. According to the force-mass-acceleration relationship, the acceleration of the cart is 0.5 m/s^2.

Let us proceed by presenting a step-by-step worked example for further illustration.” (Insert one step-by-step worked example)

Law 3: “The third law is known as the action-reaction principle. It states that for every action, there is an equal and opposite reaction. This means that if you push on an object, the object will push back on you with the same force. Let’s take a look at an example. Imagine you’re riding a bike and you hit a pothole. You’re pushing down on the front wheel, and the road is pushing back on the wheel with an equal and opposite force. This is how the action-reaction principle works.

Let us proceed by presenting a step-by-step worked example for further illustration." (Insert one step-by-step worked example)

Closing: "That's it for this video on Newton's laws. You now have a solid understanding of the law of inertia, the force-mass-acceleration relationship, and the action-reaction principle. Remember, these laws apply to the world around us and help us explain the motion of objects. If you have any questions, don't hesitate to reach out for help. Thank you for watching."

Appendix 12 Moderate guidance description

Opening: “Hello and welcome to this video lecture on Newton’s laws. My name is [Name] and I’ll be providing an overview of the three laws that explain the motion of objects. By the end of this video, you will have a general understanding of how these laws work and how they apply to the world around us. However, it’s important that you take the time to research and explore the topic on your own for a deeper understanding. Let’s get started.”

Law 1: “The first law is known as the law of inertia. It states that an object at rest will remain at rest, and an object in motion will remain in motion at a constant velocity unless acted upon by an unbalanced force. Let’s take a look at an example. Imagine a hockey puck on a frozen lake. The hockey puck is at rest and there are no forces acting upon it. According to the law of inertia, the hockey puck will remain at rest. Now, imagine a player hits the hockey puck with a stick. This action is an unbalanced force, and the hockey puck will start to move. This is how the law of inertia works.”

Law 2: “The second law is known as the force-mass-acceleration relationship. It states that force is equal to mass multiplied by acceleration. In other words, $F = ma$. This means that if you want to accelerate an object, you need to apply a force. The greater the mass of the object, the more force is needed to accelerate it. Let’s take a look at an example. Imagine you’re pushing a shopping cart. The cart has a mass of 20 kg and you’re applying a force of 10 N. According to the force-mass-acceleration relationship, the acceleration of the cart is 0.5 m/s^2.”

Law 3: “The third law is known as the action-reaction principle. It states that for every action, there is an equal and opposite reaction. This means that if you push on an object, the object will push back on you with the same force. Let’s take a look at an example. Imagine you’re riding a bike and you hit a pothole. You’re pushing down on the front wheel, and the road is pushing back on the wheel with an equal and opposite force. This is how the action-reaction principle works.”

Closing: “That’s it for this video on Newton’s laws. You now have a general understanding of the law of inertia, the force-mass-acceleration relationship, and the action-reaction principle. Remember, these laws apply to the world around us and help us explain the motion of objects. It’s important that you take the time to research and explore the topic on your own for a deeper

understanding. If you have any questions, don’t hesitate to reach out for help. Thank you for watching.”

Appendix 13 Minimal guidance description

Opening: “Welcome to this video lecture on Newton’s laws. My name is [Name], and I’ll be providing a brief overview of the three laws that explain the motion of objects. It’s important to note that this video will not provide all the information and it’s up to you to research and explore the topic on your own for a deeper understanding. Let’s get started.”

Law 1: “The first law is known as the law of inertia. It states that an object at rest will remain at rest, and an object in motion will remain in motion at a constant velocity unless acted upon by an unbalanced force. This can be demonstrated by a hockey puck on a frozen lake. If there are no forces acting upon it, it will remain at rest.”

Law 2: “The second law is known as the force-mass-acceleration relationship. It states that force is equal to mass multiplied by acceleration. In other words, F = ma. This means that if you want to accelerate an object, you need to apply a force. The greater the mass of the object, the more force is needed to accelerate it.”

Law 3: “The third law is known as the action-reaction principle. It states that for every action, there is an equal and opposite reaction. This means that if you push on an object, the object will push back on you with the same force.”

Closing: “That’s it for this video on Newton’s laws. As you can see, these laws have a lot to do with the motion of objects and how they’re affected by forces. However, this video only provided a brief overview of each law. It’s up to you to research and explore the topic on your own for a deeper understanding. I encourage you to use the resources provided below for further learning. Thank you for watching.” (Insert list of resources.)

Appendix 14 Initiate problems

Q1. Why do objects in motion tend to stay in motion, and objects at rest tend to stay at rest?

Q2. How do forces affect the motion of an object?

Q3. Why do two objects that collide always experience an equal and opposite force?

Q4. How do we explain the action-reaction principle in everyday examples?

Q5. Why does a hammer hit a nail harder than a feather?

Appendix 15 Testing questions

1. Which of the following statements best describes Newton's First Law of Motion?

- a) An object in motion will remain in motion, and an object at rest will remain at rest, unless acted upon by an external force.
- b) Force equals mass times acceleration.
- c) For every action, there is an equal and opposite reaction.
- d) The gravitational force between two objects is proportional to their masses and inversely proportional to the square of the distance between them.

2. Newton's Second Law of Motion is represented by which equation?

- a) $F = ma$
- b) $F = mv$
- c) $F = m/a$
- d) $F = a/m$

3. Which of the following statements best describes Newton's Third Law of Motion?

- a) An object in motion will remain in motion, and an object at rest will remain at rest, unless acted upon by an external force.
- b) Force equals mass times acceleration.
- c) For every action, there is an equal and opposite reaction.
- d) The gravitational force between two objects is proportional to their masses and inversely proportional to the square of the distance between them.

4. Which of the following is an example of Newton's First Law of Motion?

- a) A soccer ball rolling on the grass eventually comes to a stop.
- b) A car accelerates when the driver steps on the gas pedal.
- c) A rocket launches into space due to the force of the exhaust gases.
- d) A person pushing a stalled car on a flat road.

5. A 2 kg object is accelerating at 4 m/s^2. What is the net force acting on the object?

- a) 0.5 N
- b) 2 N
- c) 4 N

- d) 8 N

6. If you push a heavy box across the floor and it doesn't move, which of Newton's Laws is being demonstrated?

- a) Newton's First Law
- b) Newton's Second Law
- c) Newton's Third Law
- d) None of the above

7. What is the SI unit for force?

- a) Kilogram kg
- b) Newton N
- c) Joule J
- d) Watt W

8. Which of the following scenarios demonstrates an object in equilibrium?

- a) A car accelerating on a straight road.
- b) A person pushing a lawn mower at a constant speed.
- c) A ball falling freely under the influence of gravity.
- d) A book resting on a table.

9. When a person jumps off a diving board, which of Newton's Laws describes the force exerted by the person on the diving board?

- a) Newton's First Law
- b) Newton's Second Law
- c) Newton's Third Law
- d) None of the above

10. What does the term "inertia" refer to?

- a) The tendency of an object to resist a change in its motion.
- b) The force required to move an object.
- c) The gravitational force between two objects.
- d) The energy required to change an object's state of motion.

11. According to Newton’s Second Law, if the mass of an object doubles while the force remains constant, what happens to the acceleration?

- a) It doubles.
- b) It remains the same.
- c) It is halved.
- d) It is quadrupled.

12. An object is in free fall near the Earth’s surface. Which of the following best describes the forces acting on the object?

- a) Only the gravitational force is acting on the object.
- b) The gravitational force and air resistance are acting on the object.
- c) The gravitational force, air resistance, and normal force are acting on the object.
- d) No forces are acting on the object.

13. Which of the following is an example of Newton’s Third Law of Motion?

- a) A person accidentally stepping on a tack.
- b) An ice skater gliding across the ice at a constant speed.
- c) A ball bouncing off a wall.
- d) A car speeding up when the driver steps on the gas pedal.

14. If an object is moving in a circular path at a constant speed, which of Newton’s Laws explains why the object is still accelerating?

- a) Newton’s First Law
- b) Newton’s Second Law
- c) Newton’s Third Law
- d) None of the above

15. What is the net force acting on a 5 kg object moving at a constant velocity of 10 m/s?

- a) 5 N
- b) 0 N
- c) 10 N
- d) 50 N

16. In the absence of air resistance, which of the following will fall to the ground first when dropped from the same height?

- a) A 1 kg mass
- b) A 5 kg mass
- c) Both will fall at the same time
- d) It depends on their shape

17. What provides the centripetal force required for a satellite to stay in orbit around Earth?

- a) The satellite's speed
- b) The satellite's mass
- c) The force of gravity
- d) The satellite's altitude

18. Which of the following actions demonstrates Newton's Third Law?

- a) A person applying the brakes on a bike to slow down.
- b) A swimmer pushing off the pool wall to move forward.
- c) A book sliding across a table eventually coming to a stop.
- d) A baseball flying through the air after being hit by a bat.

19. If a net force of 10 N is applied to a 5 kg object, what is the object's acceleration?

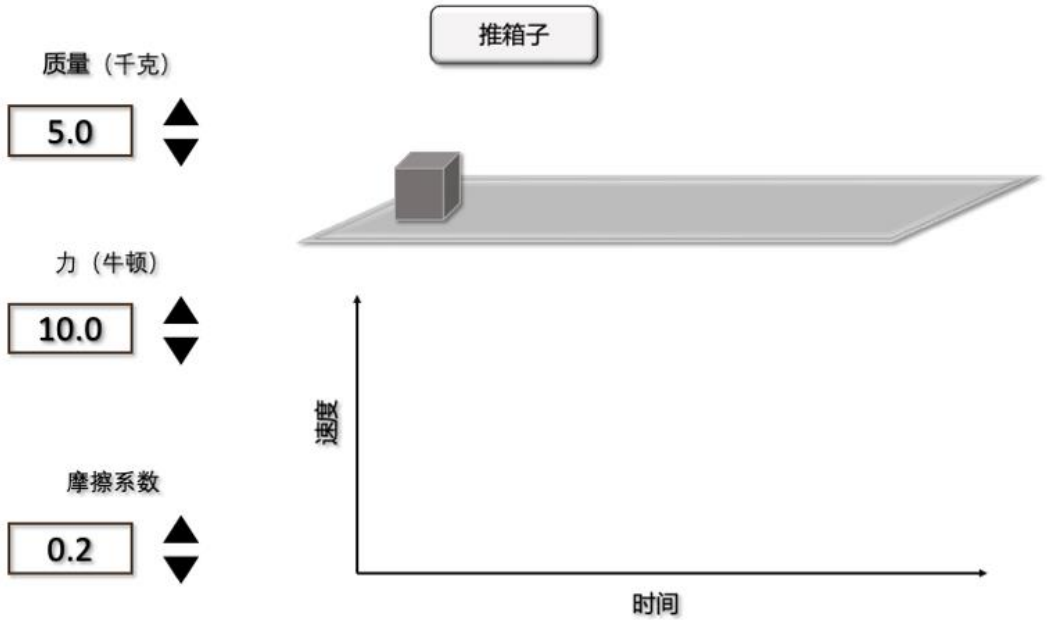
- a) 0.5 m/s^2
- b) 2 m/s^2
- c) 5 m/s^2
- d) 10 m/s^2

20. A car is moving with a constant velocity on a flat road. Which of the following is true about the forces acting on the car?

- a) The net force acting on the car is zero.
- b) The net force acting on the car is equal to the gravitational force.
- c) The net force acting on the car is equal to the normal force.
- d) The net force acting on the car is equal to the frictional force.

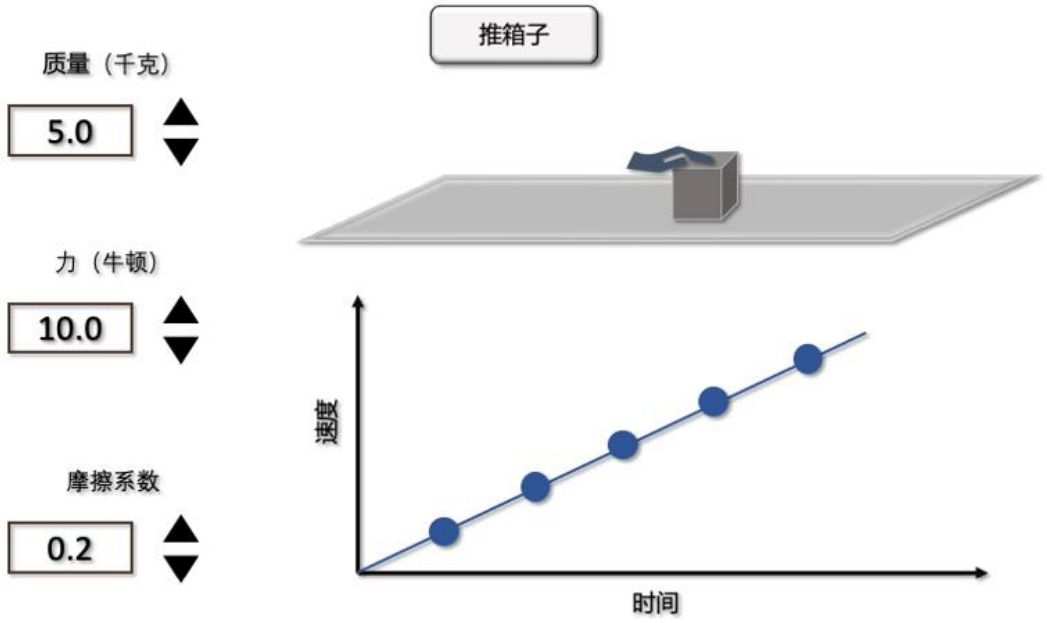
Appendix 16 Digital assistance

Figure Appendix 16.1: Illustrating Newton's Law of Motion through Interactive Software (A)



Source: Screen shot from Rainer software

Figure Appendix 16.2: Illustrating Newton's Law of Motion through Interactive Software (B)



Source: Screen shot from Rainer software

Figures Appendix 16.1 and Appendix 16.2 present an example of the interactive digital environment offered by the Rainer software for students to interpret Newton's law of motion. Through this environment, students are able to select various parameters, including the mass of a box, the force used to push the box, and the coefficient of friction. Upon pressing the button, students can visually apply the selected force to the box and observe its motion and velocity trajectory over time.

Beijing Rainer Software Technology Co., Ltd.[31] is a leading provider of professional virtual reality (VR) and interactive simulation teaching applications for primary and secondary schools.

[31] Additional information can be found on the company's official website, http://www.rainiersoft.cn/.

Appendix 17 R packages

In order to conduct the statistical analyses presented in this thesis, the open-source programming language and software environment for statistical computing and graphics, R, was utilized. R has gained increasing popularity in recent studies (e.g., D. Jiang and Kalyuga (2020)). In particular, all statistical analyses were carried out using version 3.6.2 of R. A comprehensive list of the R packages and commands utilized in this study can be found in Table Appendix 17.1.

Table Appendix 17.1: R packages and commands utilized in this thesis

Statistic analysis technique	R package	R command	Parameters/Notes
ANCOVA	stats	aov	Interest of variable should be placed in the end of the formula
ANOVA	stats	aov	
CFA	lavaan	cfa	std.lv = Ture
Cronbach's Alpha	psych	alpha	
Generating statistical graphics	ggplot2	ggplot	se = "bootstrap"
LMM	lme4	lmer	The random term is input as (1 | class); REML = True
MANCOVA	stats	manova	Pillai's statistics to get approximate F value; Interest of variable should be placed in the end of the formula
MANOVA	stats	manova	Using Pillai's statistics to get approximate F value
McDonald's Omega	psych	omega	
OLS	stats	lm	
Path analysis	lavaan	sem	se = "bootstrap"
Tukey's range test	stats	TukeyHSD	

Source: Compiled by author

Table Appendix 17.1 presents different statistical analysis techniques and the corresponding R packages and commands used to perform them in this study. Each row lists the name of a statistical technique, the R package that contains the technique, the R command to execute the analysis, and some additional parameters or notes relevant to the use of the command. For

example, ANCOVA analysis is performed using the `aov` command from the `stats` package, and the interest of variable should be placed at the end of the formula. Similarly, CFA is performed using the `cfa` command from the `lavaan` package with the parameter std.lv = True, and Tukey's range test is performed using the `TukeyHSD` command from the `stats` package.

Appendix 18 Bootstrapping for estimating standard errors in path analysis

The standard errors of the path analyses performed in the present thesis were calculated through the utilization of the within-sample bootstrapping method. This method involves resampling the original sample repeatedly to create multiple datasets, each with the same sample size as the original dataset. This is performed with the intention of estimating the variability of the path coefficients by conducting path analysis on each of the resampled datasets.

Suppose the original dataset is represented as vector Y, and the aim is to estimate the standard errors of the path coefficients in the path analysis model. The following steps outline the procedure for using the within-sample bootstrapping technique:

1. Draw a sample of size n from the original dataset Y, represented as Y^*. This sample will be used to estimate the parameters of the path analysis model.
2. Conduct path analysis on Y^* and estimate the parameters of the path analysis model.
3. Repeat steps 1 and 2 B times, where B is a large number (such as 1000 or 5000), to generate multiple samples and estimates of the parameters of the path analysis model.
4. The standard error of each path coefficient in the path analysis model can be calculated by taking the standard deviation of the B estimates of each parameter.

This procedure provides a more precise estimate of the variability of the path coefficients, enabling more informed inferences to be made regarding the relationships between the variables in the model. Mathematically, the standard error of a path coefficient (β) can be expressed as:

$$SE_\beta = \sqrt{var(\hat{\beta})}$$

where $\hat{\beta}$ represents the estimated value of β and $var(\hat{\beta})$ represents the variance of $\hat{\beta}$ across the B iterations of the bootstrapping procedure.

In the context of the present thesis, the within-sample bootstrapping technique can be utilized to estimate the standard errors of the path coefficients in the path analysis model, providing a more accurate estimate of the variability of the path coefficients and allowing for more informed inferences regarding the relationships between the variables in the model.