Test preparation: An integrated thesis of a meta-analysis and studies on students' strategies



Zhanxin Hao

Department of Education

St Antony's College

A thesis submitted for the degree of

Doctor of philosophy in Education

Hilary 2022

Abstract

This integrated thesis investigates the effect of test preparation in educational contexts as well as students' usage of strategies when they prepare for tests. One meta-analysis and four (quasi-) experimental studies were included in the thesis, constituting three research papers. In addition to the main findings reported in the three papers, some cross-experimental analyses were added to further contribute to the overall interpretation of research findings.

The meta-analysis of test preparation effects on large-scale educational tests suggested that students' test performance can be significantly improved by test preparation. The breadth of test preparation intervention, workbooks, training test-taking skills and test preparation management strategies were found to be significant moderators of test preparation effect. However, the meta-analysis found little evidence to support the notion that testing (i.e., the use of sample items and practice tests) has a superior impact on test performance over other practices (i.e., lecture or restudy), contrary to the solid testing effect reported in the literature.

The effects of testing and restudy strategies for test preparation were further explored across four (quasi-) experiments, which were conducted in an educational psychology course in a teacher training college in China. The first three (quasi-) experiments manipulated three variables: test preparation strategies, the stakes of the test preparation setting, and metacognitive training and practices. How these variables affected students' performance on post-tests and delayed course tests were investigated. Under high-stakes settings, testing with feedback had significant effects on both post-tests and delayed course tests compared with the control group while restudy only outperformed the control on post-tests. Embedding metacognitive training and monitoring and control tasks during test preparation did not significantly boost the effects. The fourth quasi-experiment simulated a more realistic test preparation condition where students could choose either testing or restudy to study the materials multiple times. Students tended to use one-time testing more often, and they were reluctant to restudy a topic repeatedly. Both quantitative and qualitative analyses suggested that students' strategy uses were predicted by their confidence and the question type. Chinese students in this research could recognise both the direct and indirect effects of testing practice, which seems to be culturally distinctive. However, additional analyses revealed that students' blended use of the two strategies failed to enlarge the test preparation effect. Across the four experiments, students with lower prior attainment seem to benefit more from the test preparation interventions.

Acknowledgement

Doing this DPhil is a challenging but enjoyable journey, with moments of both joy and loss. This journey would not have been possible without the support and companionship of many people.

First and foremost, I would like to express my deepest gratitude to my two supervisors, Professor Jo-Anne Baird and Dr Yasmine El Masri. Every supervision meeting we had was constructive and motivating. Their impeccable knowledge and expertise in the field have guided me to complete the work that I could not have imagined when I started the DPhil. Their constant support and care for both my DPhil and my career development are invaluable. They are my life role models, inspiring me to grow into a female researcher as wise, independent, caring, and internally strong as they are.

I would like to thank my examining committees, Professor Michelle Meadows, Professor Sonali Nag, Professor Therese Hopfenbeck, and Dr Kit Double, for their valuable feedback during my Transfer of Status and Confirmation of Status interviews.

I would also like to express my gratitude to all participants who were willing to participate in this research. I am particularly grateful to Mr. Lijiang for his kindness and in helping me contact the college and organise the implementation of data collection. It is a miracle that this was done during the pandemic. His cooperation and assistance made this research possible.

I am thankful for my dear friends who always encourage me to face challenges with courage and a positive mindset. The very joyful and relaxed times we spent together at Oxford are my precious memories.

Last, my heartful thanks go to my family. I am grateful to my parents, grandparents, and all my family members for their unconditional love, understanding, and support. I would like to thank Zhaojun, my beloved boyfriend, whose love, support, and patience have been my anchor to navigate all the ups and downs during this journey.

Table of Contents

Abstract ..... **ii**

Acknowledgement ..... **iii**

Table of Contents ..... **iv**

List of Tables and Figures ..... **viii**

Abbreviations ..... **xi**

Part I: Background to the Research **1**

Chapter 1. Introduction 2

1.1 Educational tests and test preparation 3

1.2 The current research 6

1.3 Structure of the thesis 7

Chapter 2. Literature Review 11

2.1 The definition of test preparation 11

2.2 Empirical studies on test preparation practices 21

2.3 Effects of test preparation 27

2.4 Theoretical underpinnings for the test preparation effect 38

2.5 Summary 43

Chapter 3. Methodological Strategy 46

3.1 Rationale of the research 46

3.2 Overall methodology of the four experimental studies 56

3.3 Ethical considerations 78

Part II: Research Papers **80**

Chapter 4. Paper 1 81

Abstract 83

1. Introduction 84

2. Evaluating the evidence of test preparation 85

3. Method 95

4. Results 107

5. Discussion 125

6. Limitations 130

7. Conclusion 132

References 132

Author Contributions 138

Chapter 5. Paper 2 139

Abstract 141

Introduction 142

Experiment 1: The Effect of Testing and Restudying During Preparation for a High Stakes Test 149

Method 149

Results 154

Summary of Experiment 1 156

Experiment 2: Testing and restudying effect in a low-stakes context 157

Method 158

Results 159

Summary of Experiment 2 161

Experiment 3 The Effect of Using Metacognitive Strategies with Restudy and Testing for Test Preparation 162

Method 163
Results 168
General discussion 176
Limitations 181
Conclusions 182
References 183
Author Contributions 189
Chapter 6. Paper 3 190
Abstract 192
1. Introduction 193
2. Method 198
3. Results 206
4. General Discussion 220
5. Conclusion 226
References 227
Author Contributions 236
Part III: Synthesis 237
Chapter 7. Overview of findings and additional analysis 238
7.1 Summary of findings of the three papers 238
7.2 The effect of students' combined use of strategies in Experiment 4 243
7.3 Test preparation effect across experiments 247
7.4 Prior attainment levels and score gains 250
7.5 Prior attainment levels, test preparation behaviours and monitoring accuracy 265
7.6 Overview and discussion of research findings 275
Chapter 8. General Discussion 280

8.1 Discussion of the overarching research questions 280
8.2 Implications 298
8.3 Limitations and Future directions 301
8.4 Concluding remarks 304

References 306

Appendices 328

Appendix 1: Ethics Approval 328
Appendix 2: Information Sheet 329
Appendix 3: Test papers (exemplar) 333
Appendix 4: Answer sheets (exemplar) 335

List of Tables and Figures

Table 1 Information of participants in four experiments. 58

Table 2 Design of Experiment 3. 60

Table 3 Examples of practice questions and statements. 63

Table 4 The structure of the pre/post-tests used in four (quasi-) experiments. 66

Table 5 Inclusion and exclusion criteria. 97

Table 6 Descriptive characteristics of the included studies. 108

Table 7 Results of the sub-group analysis. 113

Table 8 Results of meta-regressions. 116

Table 9 Experimental and Control conditions in Experiment 1. 149

Table 10 Descriptive statistics of pre-test, post-test and final course test in Experiment 1. . 154

Table 11 Descriptive statistics of course test, pre-test, and post-test scores in Experiment 2. 159

Table 12 Experimental and Control conditions in Experiment 3. 164

Table 13 Study behaviours of students in experimental groups in Experiment 3. 168

Table 14 Mixed effect model results of the learning attempt. 171

Table 15. Descriptive statistics of pre-test, post-test, and course test scores in Experiment 3. 172

Table 16 Crosstabulation of Study Attempt and Strategy Use. 206

Table 17 The frequency of different strategy use patterns. 208

Table 18 Students' average metacognitive ratings on each topic. 209

Table 19 GLMMs results of the first study attempt (Model 1), the repeated study decision (Model 2), and the second study attempt (Model 3). 212

Table 20 Themes and codes of students' perceptions on when to use rereading. 216

Table 21 Themes and codes of students' perceptions on when to use testing. 219

Table 22 Test performance on pre- and post-tests and course test in Experiment 3 and 4. . 244

Table 23 Results of moderator analysis of the CCMA. 249

Table 24 Score gains among absolute prior attainment levels in the four (quasi-) experiments. 252

Table 25 Score gains among relative prior attainment levels across test preparation conditions in each experiment. 259

Table 26 Significant difference in study attempts between test preparation conditions within prior attainment levels 269

Table 27 Absolute monitoring accuracy of the pre-test, post-test and changes. 272

Figure 1 The whole process and components of test preparation. 20

Figure 2 Structure of Cognitive Loads (Sweller, 1994, 2010). 40

Figure 3 An overview of four (quasi-) experiments in the research. 51

Figure 4 The design of Experiment 1 and Experiment 2. 53

Figure 5 The design of Experiment 3 and Experiment 4. 54

Figure 6 The Intervention design of Experiment 3. 71

Figure 7 The intervention design of Experiment 4. 72

Figure 8 Flow diagram of literature search and processing of records. 99

Figure 9 Forest plot of all recorded effect sizes. 112

Figure 10 A funnel plot showing the relationship between standard error and observed effect size for the meta-analysis. 125

Figure 11 The interaction between group and confidence on study attempts. 172

Figure 12 Participants' strategy choices among their first two study attempts. 207

Figure 13 Patterns of students' strategy uses and metacognitive ratings. 210

Figure 14 Significant and non-significant moderators of test preparation effects in the meta-analysis in Paper 1. 239

Figure 15 Significant differences from ANCOVAs on post-test performance between test preparation conditions across Experiments 1 to 3. 240

Figure 16 Significant differences from ANCOVAs on course-test performance. 241

Figure 17 Significant differences from ANCOVAs on the post-test and the course-test performance across test preparation conditions in Experiment 3 and 4. 245

Figure 18 Forest plot of effect sizes of all test preparation interventions in the four (quasi-) experiments. 248

Figure 19 Block diagram of significant differences between different prior attainment levels on score gains in Experiments 1 and 2. 254

Figure 20 Block diagram of significant differences between different prior attainment levels on score gains in Experiments 3 and 4. 256

Figure 21 Block diagram of significant group differences between different relative prior attainment levels on score gains on post-tests. 262

Figure 22 Block diagram of significant group differences between different relative prior attainment levels on score gains on final course tests. 264

Figure 23 Block diagram of significant group differences between different prior attainment levels and test preparation conditions on time use during the intervention. 268

Figure 24 Significant differences between prior attainment levels and test preparation conditions on study attempts. 270

Figure 25 Significant differences between prior attainment levels and test preparation conditions on testing and restudy attempts. 271

Abbreviations

ACT	The American College Testing
ANOVA	The analysis of variance
ANCOVA	The analysis of covariance
CCMA	Continuously Cumulating Meta-Analysis
CET4	College English Test Band 4
CLEP	College Level Examination Program
CLT	Cognitive Load theory
EAP	English for Academic Purposes
EFL	English as a Foreign Language
FTE	Free Teacher Education
GLMM	General linear mixed-effects model
IELTS	International English Language Testing System
K-12	Kindergarten through grade 12
KLI	Knowledge-learning-instruction
MCQ	Multiple-Choice Questions
NTCE	National Teacher Certification Examination (in China)
PISA	Programme for International Student Assessment
PSAT	Preliminary Scholastic Assessment Test
RQ	Research Question
RVE	Robust Variance Estimation
SAT	The Scholastic Aptitude Test
SD	Standard Deviation
SE	Standard Error
TAP	Transfer-Appropriate Processing
TOEFL iBT	Test of English as a Foreign language Internet-based Test
UMAT	Undergraduate Medical and Health Sciences Admission Test
U.S.	The United States
ZPD	Zone of proximal development

Part I
Background to the Research

Chapter 1. Introduction 2

Chapter 2. Literature Review 11

Chapter 3. Methodological Strategy 46

Chapter 1. Introduction

Due to the high stakes associated with students' test performance in educational contexts, test preparation has been widely employed to maximise test scores. It is not surprising that preparation strays into intensive cramming in real-world practice; for instance, teachers narrow the curriculum to test-related content, students burn the midnight oil to study for the exam, and some parents pay expensive tuition fees for private tutoring. Test preparation practices are prevalent across the world for different kinds of educational tests, both curriculum-based and non-curriculum-based tests. Despite the fact that non-curriculum-embedded or aptitude tests, such as the Scholastic Aptitude Test (SAT), are not supposed to be taught in schools, numerous commercial institutions provide pricey preparatory SAT courses and assert their efficacy (Buchmann et al., 2010). However, in academic literature, the effect of test preparation is mixed. In order to determine the worth of the time, effort and money invested into test preparation for various educational tests, it is necessary to depict the impact of different test preparation practices on different types of tests in different contexts. Besides, students' self-study for test preparation is a common practice, though receives less research attention. It is crucial to comprehend how students use different strategies in their self-study and identify conditions in which different strategies are effective, to better direct students' preparatory practices for high-stakes tests. Based on both synthetical and empirical evidence, this thesis aims to build a comprehensive understanding of how test preparation affects test performance in educational tests, as well as how students use different strategies in self-study for test preparation.

This chapter introduces the background to the current research programme, stating how test preparation has become a topic of interest for both real-world practice and research. Then the current research is outlined, including the overarching research questions and a brief

overview of all studies making up the research programme. The overall structure of the thesis is also presented in this chapter.

1.1 Educational tests and test preparation

Educational tests have been an inherent part of education systems throughout the world. Test results have been used tremendously for making decisions related to accountability, policymaking, and gatekeeping by educational authorities internationally and nationally (Newton, 2007). Some decisions may link to rewards or sanctions for different stakeholders (e.g., students, teachers, schools), making those tests high-stakes (Madaus, 1988). High-stakes tests usually place test-takers on an ability dimension according to their test scores, and their future educational or occupational status might be determined accordingly (Stobart & Eggen, 2012). The increasing power of test scores has spawned an ever-growing demand for test preparation activities inside and outside schools, aimed at maximising students' test scores.

Worldwide, countless commercial educational companies have provided students with test preparation services, including classroom-based courses, private tutoring, online courses as well as test preparation materials. These services mainly pertain to college admission tests, but also cover other tests in K-12 stages (Bray, 1999, 2017). It is claimed that these test preparation courses are able to significantly improve students' chances of successfully moving through the allocation process (Baker & LeTendre, 2005; Stevenson & Baker, 1992), and such a claim has been repeatedly sold by companies and believed by students, making the test preparation industry lucrative.

An over-reliance on test results in educational allocation processes has not only nourished the thriving commercial test preparation industries; teaching and learning activities in schools have also been affected. Some schools narrow their curriculum to tested subject,

and non-tested subjects are neglected (Au, 2007; Wyn et al., 2014). In classrooms, innovative teaching strategies give way to teacher-centred instruction which focuses on frequently tested content and students are drilled using worksheets repeatedly (Blazar & Pollard, 2017; Cho & Chan, 2020; Firestone et al., 2004). Across the world, formal schooling has been compromised by preparing students for tests to different extents. It is reported that in the U.S., more than 85% secondary schools provide college admission test preparation courses (Scholes & Lain, 1997). The situation is even more intense in countries with an examination-oriented culture, such as China and Korea (Lee & Larson, 2000; Zhang, 2011). For example, in China it is not uncommon for high schools to cover the required national curriculum within the first two years, so that the final year can focus exclusively on test preparation by repeating tested content as well as drilling students' test-taking skills (Kwon et al., 2017; Ye, 2015). A lot of cramming schools (college-preparation schools) even supplant the ordinary curriculum with a curriculum specifically designed around test preparation activities (Byun, 2014; Kwok, 2010; Wang, 2022).

The growing prevalence of test preparation raises a lot of questions. Educational companies and cramming schools have advertised significant impacts of their test preparation courses on test score gains, while the tuition fees of these commercial courses can be substantial, thus increasing the financial pressures on students and their families and affecting the equal access for all prospective students (Yu & Green, 2021). Furthermore, preparing for a test through either attending courses or self-study tends to take a great deal of time and energy, leaving less time for participation in other learning activities. Given the significant amount of time and expense required by test preparation activities, as well as the potential detrimental impact on school curriculum, the fundamental questions are:

- Does test preparation work, and to what extent does test preparation work?
- What test preparation activities or strategies work better?

Although existing research has tried to address these questions, the effect of test preparation remains a controversial subject.

The magnitude of the effect across studies varies widely. For example, some studies confirmed a positive effect and concluded that test preparation could motivate students and increase test scores (e.g., Appelrouth et al., 2017; Briggs, 2009; Bunting & Mooney, 2001; Kulik et al., 1984; Messick, 1982). By contrast, a number of studies asserted that preparation has no effect, or the score gain is extremely small (e.g., Li & Xiong, 2018; McGaghie et al., 2004; Powers & Rock, 1999). As test preparation can be a complicated learning process driven by different participants (e.g., teacher-driven coaching; student-driven self-study), using various materials and strategies, and adherence to various timetables. The intrinsic variation of test preparation activities may result in variability in findings in test preparation effect studies. Researchers have remarked upon the need to identify possible elements of test preparation activities. Becker (1990) suggested that to gain a clear understanding of the effect of test preparation, it is important to determine its constitution. By linking exactly what students do during test preparation and the key features of their practice to the effect on test performance, we can gain a clearer view of how to develop optimal test preparation activities to improve test results and students' long-term learning (Arendasy et al., 2016; Montgomery & Lilly, 2012).

It should be noted that the current literature on test preparation is primarily focused on teacher-led activities (i.e., coaching), leaving students' self-study for the test under explored. However, in some cases (e.g., in higher education settings), preparation practices are self-paced and self-directed. Therefore, it is important to explore the relationship between features of students' test preparation practices and the corresponding score gains, to better instruct students' self-study. Besides, most of the research on test preparation effect has been conducted in West. Although there may be similarities between test preparation practices and

experiences across different cultural contexts, there are also likely to be important differences that are specific to countries with different educational practices, assessment systems, as well as cultural and societal expectations on assessment outcomes (Jeynes, 2008; Leung, 2006). It would be valuable to investigate whether test preparation practices and effects vary in different cultural contexts, which would also inform the development of more effective and culturally appropriate approaches to test preparation, as well as contribute to a broader understanding of the impact of high-stakes testing on student learning around the world.

1.2 The current research

This research investigates test preparation effects in educational settings using both synthetical and empirical evidence from a series of interrelated studies. Note that empirical evidence in this research was collected from China, where educational systems and assessment cultures are different to the West (Davey et al., 2007; Niu, 2007). This research can extend the literature and could have important implications for policy and practice in China's context.

The overarching aim of this research is to understand the overall effect of test preparation on students' performances on large-scale educational tests, as well as how students regulate their utilisation of different strategies for test preparation. The current research is a meta-analysis and four interrelated (quasi-) experimental studies. Specifically, the first study is a meta-analysis which investigates the overall estimate effect of test preparation interventions on large-scale educational tests by synthesizing previous literature. The moderating effects of possible characteristics on the direction and magnitude of the test preparation effect have also been analysed.

Building upon the findings of the meta-analysis, a series of subsequent four experimental studies was designed to compare the effects of two common test preparation

strategies—testing and restudy—in ecologically valid educational settings in China, to explore the conditions under which test preparation effect can be maximised. Additionally, how Chinese students regulated their use of the two strategies, factors that may relate to students' choices, and their perceptions of the effectiveness of the two strategies are examined, to advance the interpretations of students' study behaviours in test preparation contexts.

In general, the present research aims to answer the following research questions (RQ):

RQ1. What effect do test preparation interventions have on test performance?

RQ2. What factors affect the effectiveness of test preparation interventions?

RQ3. How do Chinese students regulate their use of testing and restudy? What factors influence their strategy choices?

1.3 Structure of the thesis

This integrated thesis is presented in three parts. This, first part, is the background of the research programme, consisting of three chapters: Introduction, Literature Review, and Overall Methodology. The second part of the thesis is composed of three self-standing papers, which report the main findings. The third part integrates the thesis and is composed of two chapters: Overview of Findings and Additional Analyses, and General Discussion. This part summaries and integrates findings across the thesis, interprets relevant findings from a synthesised perspective, and provides a general discussion around overarching research questions. The implications and possible limitations of this research are discussed in the third part, as well as directions for future research. Given the complex nature of the studies and the integrated structure of the thesis, a detailed description of each subsequent chapter follows, to guide the reader.

Chapter 2 redefines the term of 'test preparation' based on a holistic perspective and reviews previous studies on test preparation. Given that this integrated thesis consists of three research papers, each of which includes a section on literature review. To avoid repetition, the literature that quantified test preparation effects through (quasi-) experiments, and studies related to testing/restudying effect, as well as literature on students' regulation of testing practices are reviewed in individual papers. Chapter 2 aims to provide the theoretical and empirical foundations to guide the overall research design and mainly reviews literature that is not covered in the manuscripts.

Chapter 3 proposes the overall research design of this integrated thesis and specific methodological strategies adopted by each individual study. The analytical approach applied in the three papers and in the additional analysis are also described. This chapter ends with a statement of ethical considerations.

Chapter 4 presents the first research paper (Paper 1) of the meta-analysis of test preparation effects on educational tests. This meta-analysis mainly concerned the effect on large-scale educational tests, as the main test preparation literature is developed based on discussions around the large-scale tests, most notably college admission tests and English tests. Thus, relevant data are more substantial and accessible. Possible moderators of test preparation effect were discussed based on the sub-group analysis and meta-regression results in this paper. Findings of this paper sparked research interests in the effects of using testing and restudy as test preparation strategies in students' self-study, which were explored in subsequent (quasi-) experiments.

Chapter 5 displays the second research paper (Paper 2), which investigated the effects of using testing and restudy during test preparation sessions on students' performance of educational psychology tests. Whether effects of the two strategies varied across different

conditions had been probed in this paper. This paper reports findings from three (quasi-) experiments.

Chapter 6 presents the third research paper (Paper 3), which probed how students regulated their uses of testing and restudy, factors associated with their strategy choices, as well as their perceptions on the efficacy of the two strategies. Students' preferences between the two strategies were discussed from a sociocultural perspective in this paper.

Chapter 7 consists of additional analyses integrating data across the four (quasi-) experiments. This chapter first summarises findings reported in the three papers. Given that Paper 2 and Paper 3 are written in a narrative structure with a specific focus in each paper, only relevant result collected from the (quasi-) experiments are discussed coherently to build the narratives. Some data were not central to the narrative thread and thus did not fit in the two individual papers. These data might also provide important insights or alternative interpretations for the results. Thus, these 'overlooked' data were analysed and reported in Chapter 7. The first additional analysis investigated the effects of students' combined use of testing and restudy on their performance of the post-test and the delayed course test and compared the effect of combined use of strategies with the effect of using single strategies. Then, a continuously cumulating meta-analysis (CCMA) was conducted to synthesise an overall effect of test preparation interventions across the four (quasi-) experiments. The CCMA provides a synthetical estimate of the effect of test preparation on small-scale, course tests, which could be regarded as an important complement to the meta-analysis on the large-scale tests in Paper 1. Moreover, how individuals with different prior attainment levels use testing and restudying as well as how they benefit from testing and restudy practices to various degrees was analysed and presented in this chapter.

Chapter 8 is the final chapter, which integrates and discusses main findings from all included studies. The implications of this thesis, both on the theoretical development and

real-world practice, have been discussed. The limitations of this research are outlined, indicating the directions for future research. At the end of the chapter, and the end of the thesis, an overall summary of key findings and concluding remarks are provided.

Chapter 2. Literature Review

This chapter reviews the literature on test preparation practices and effects to provide theoretical and empirical foundations to inform the overall design of the research. The first section provides the definition of test preparation and proposes a framework of the whole test preparation process based on the literature. The second section reviews test preparation practices employed by teachers and students as well as the reasons for their choice of specific practices. Later, the effect of test preparation is reviewed in terms of different influential factors. The fourth section outlines several theoretical frameworks that are helpful to explain the effect of test preparation. Reflections on the literature and limitations of previous research are presented at the end of each section. Finally, motivation behind the research in the thesis is briefly outlined, in preparation for presentation of the overall methodology in the next chapter.

2.1 The definition of test preparation

Although the general concept of test preparation is well understood within the field of education and assessment, there are nuances and variations in how it is defined in the literature. Various terms have been used to refer to relevant practices and activities. However, due to the loose use of various terms, there might be ambiguity about their precise meanings in different contexts. Therefore, clarification of relevant terms is necessary to avoid misunderstanding and ensure consistency in the use of terminology in this research. The following subsections clarify the terms used in the literature and present useful frameworks describing practices related to test preparation. Based on the understanding of the literature, a framework for test preparation processes and components is summarised in this research.

2.1.1 Terms that relate to test preparation

In the literature of educational tests and test preparation, various terms are used to refer to test preparation behaviours, including 'teaching to the test', 'coaching', 'shadow education', 'teaching test-taking skills', 'test wiseness', 'test familiarisation'. These terms are often used to describe either a single test preparation procedure or a more elaborate test preparation process that encompasses multiple behaviours. Here, the meanings of these terms are differentiated and the contexts in which they are used frequently is discussed.

Coaching

Coaching is one of the most recognizable and representative terms in the test preparation literature. It refers to training programmes structured by experienced teachers or commercial companies, which is often additional to formal schooling, but can also be courses provided by schools. Led by teachers, coaching usually targets specific high-stakes tests which assess students for educational selection or certification. For example, in the U.S., many companies provide coaching on college admission tests – the SAT and the American College Testing (ACT) (Briggs, 2009; Buchmann et al., 2010). Coaching has been regarded as a broad concept, ranging from short-term training on test content to long-term instruction on general subject knowledge development (Messick, 1982; Scholes & Lain, 1997). Anastasi (1981) described three levels of coaching: 'a) short test-taking orientation sessions; b) longer coaching programmes that include intensive, concentrated drill or 'cramming' on sample test questions; c) instruction in broad cognitive skills' (p.151). In practice, the first two levels are most commonly observed, and thus this term tends to narrowly indicate intensive interventions on specific test taking (Messick, 1982).

Shadow education

Shadow education overlaps with coaching to some extent. It refers to students attending commercial courses or having private tutoring outside of the formal education system

(Allensworth et al., 2008; Bray, 1999). Thus, all coaching activities outside of formal school settings can be recognised as shadow education. However, shadow education has a relatively broader meaning than coaching, as preparing students for tests is not its only purpose. Shadow education can also be used to provide students with enrichment in subjects that they want to excel in or develop knowledge and skills that are not covered in the regular school curriculum.

Teaching to the test

The term teaching to the test describes the teaching practices that restrict instruction to tested content and focus classroom practice on questions that are identical or very similar to test items (Miyasaka, 2000; Popham, 2001). One of the important initiatives of teaching to the test is to improve students' performance on local or national achievement tests to help schools reach the accountability requirements (Au, 2007; H. Li & Xiong, 2018). It is usually a manifestation of washback of high-stakes testing on curriculum and teaching, which has been identified in studies across countries (e.g., Andrews et al., 2002; Cheng, 2005; Popham, 1991) and across tests contexts, such as K-12 contexts (e.g., Crocker, 2003) and language tests (e.g., Read & Hayes, 2003).

Test familiarisation

The term test familiarisation refers to using methods to familiarise test-takers with the mechanics of test taking (Arendasy et al., 2016). Previous literature highlighted several commonly used test familiarisation methods, including using workbooks, practice sample tests and online sources (e.g., Arendasy et al., 2016; Messick, 1982). Through these methods, students could gain information on test administration, test formats, sample items, exemplar answers and marking criteria.

Test-wiseness

Test-wiseness is used interchangeably with teaching (or practicing) test-taking skills, indicating specific skills taught by teachers or used by test-takers while responding to a test, which provides them with the capacity to utilise the specific test characteristics and formats to achieve high scores (Bangert-Drowns et al., 1983; Millman et al., 1965; Popham, 2001). General test-taking skills include time management, error-avoidance, optimal guessing, pattern guessing, deductive reasoning and so on (Mehrens et al., 1998).

The terms mentioned above have been frequently used in the literature, and each of them captures multiple test preparation behaviours. In this research, test preparation is defined from a holistic and all-inclusive perspective proposed by Messick (1982):

> *Any intervention procedure specifically undertaken to improve test scores, whether by improving the skills measured by the test or by improving the skills for taking the test, or both.* (Messick, 1982, p.70).

In other words, any learning and teaching practices aiming to improve test scores could be regarded as test preparation. Based on this holistic perspective, test preparation can be regarded as a complicated endeavour which is sometimes integrated within regular teaching and learning practices. However, test preparation, especially preparation for high-stakes tests, differs from regular teaching and learning as it is usually associated with greater intensity and a narrow focus. In this regard, not all learning strategies can be considered as test preparation strategies. Some general teaching and learning strategies that allow students to learn and understand new information or develop general cognitive and noncognitive abilities, such as collaborative learning and spaced learning, may be not directly related to test preparation. It can be challenging to dissociate and untangle test preparation from teaching and learning, thus, specific test preparation and test-taking contexts, as well as the purpose of relevant

practices need to be considered to comprehend what constitutes 'test preparation'. Previous researchers proposed several frameworks to classify test preparation practices and evaluate the validity of interpretations of score gains from different test preparation practices, contributing to a comprehensive understanding of the relationship between test preparation, testing, teaching and learning. These frameworks are reviewed in the next subsection.

2.1.2 Frameworks describing test preparation

As discussed in the section above, test preparation is a form of special teaching and learning that occurs in the context of preparing for tests, particularly high-stakes tests where the results have significant consequences. The administration of tests, especially high-stakes tests, can influence teaching and learning practices, as teachers and students may have to engage in specific test preparation activities to improve test scores. Understanding the impact of tests on teaching and learning can provide insights into the different dimensions and intensity of test preparation activities. Therefore, Alderson and Wall's (1993) framework on washback effect was firstly reviewed to understand what aspects that test preparation may focus.

Alderson and Wall (1993) Framework of washback

It has been widely acknowledged that educational tests, especially high-stakes tests, have great impacts on teaching and learning, which are referred to as *washback* or *backwash* effect (Alderson & Wall, 1993; Au, 2007; Baird et al., 2017). Alderson and Wall (1993) posited fifteen possible washback hypotheses relating to the impact of a test on teaching and learning, including the content (e.g., *what* teachers teach and learners learn), methods (e.g., *how* teachers teach and learners learn), order (e.g., the *rate* and *sequence* of teaching and learning), extents (e.g., the *degree* and *depth* of teaching and learning), as well as consequences (e.g., a test with *important consequences* will have washback), teachers and students' attitudes, and different impact on different persons (e.g., tests will have washback

effects for s*ome teachers and some learners*, but not for others). These hypotheses of test washback pinpoint aspects that teachers and students' preparation for tests may focus upon, including content, methods, extents, contexts, and attitudes (e.g., student motivation). Meanwhile, this framework suggested that the consequences of the test (i.e., the stakes associated with test results) and individual differences should be taken into account when studying test preparation.

Based on the framework of washback effect of tests, researchers have extensively studied how high-stakes tests impact the instructional methods used by teachers and the learning strategies adopted by students (Andrews et al., 2002; Firestone et al., 2004; Smith, 1991; Xie, 2013). These impacts, in turn, can have significant implications for the interpretation and use of test scores (Koretz et al., 2001; Messick, 1996; Miyasaka, 2000). Some researchers described and classified test preparation process by differentiating teachers' practices and evaluate the validity of interpretation of score gains from test preparation practices. In the following, three frameworks proposed by Smith (1991), Koretz et al. (2001), and Miyasaka (2000), are reviewed.

Smith (1991) Test preparation categorisation

Based on results of qualitative interviews with teachers, school administrators, testing experts and school critics, Smith (1991) categorised a wide range of teaching practices that prepared students for external, mandated, high-stakes tests in U.S. elementary schools. Specifically, eight types of preparation for high-stakes achievement tests were summarised: (a) ordinary curriculum with no special preparation; (b) teaching test-taking skills; (c) trying to exhort pupils to do their best; (d) teaching content known to be covered by the test; (e) teaching about the test in format and content; (f) stress inoculation; (g) practicing tests or parallel test items; and (h) cheating, such as providing students with hints or even correct answers or altering marks on answer sheets. Smith highlighted that 'political' use of test

scores and the distinction between ordinary instruction and tests compelled teachers to resort to systematic test preparation to improve scores and avoid criticism, thereby affecting their autonomy to a large extent (p.540). In this work, Smith provided a reasonably comprehensive categorization that captured teachers' responses to high-stakes testing and the dilemmas they encountered within accountability-driven testing systems. This framework lays a foundation for understanding the relationship between high-stakes testing, test preparation and teaching practices.

Koretz et al. (2001) Test preparation categorisation and validity of score gains

Similar to Smith's (1991) categorisation, Koretz et al. (2001) focused on teachers' responses to external testing and classified seven types of preparation practices: (a) teaching more; (b) working harder; (c) working more effectively; (d) reallocation; (e) alignment; (f) coaching; and (g) cheating. Teaching more means taking more time on instructional activities, working harder indicates covering more materials during classes; and working more effectively suggests adopting more effective teaching methods or better curriculum. Reallocation implies shifting instructional time to focus more on test-related material, and alignment is a special form of reallocation, which means that certain content is assigned a more prominent role in the teaching and learning process than other content. Coaching refers to either a focus of test preparation on elements of the test that are entirely unrelated to the domain, or instruction emphasis on narrow aspects of substantive performance elements to match the emphasis of test items. Cheating is the most extreme form of test preparation and can be diverse, including providing answers or access to secure test items, violating test administration procedures, correcting students' responses, and so on. Koretz and his colleagues believed that the first three types of test preparation – i.e., teaching more, working harder and working more effectively – can produce meaningful gains in scores, while the others, i.e., reallocation, alignment, coaching and cheating, might cause inflated scores.

Inflated scores have significant implications for the validity of the assessments as their interpretation overstates what test takers know and can do (Kane, 2012; Messick, 1982, 1989).

Miyasaka (2000) Main areas of test preparation

Miyasaka (2000) proposed a more generalised framework to classify main areas of test preparation and evaluate the validity of score gains from different test preparation practices. Five major areas of test preparation in the context of large-scale tests were identified: (a) relationship between curriculum and test content; (b) assessment approaches and test formats, (c) test-taking strategies; (d) timing of test preparation; and (e) student motivation. Based on previous studies, Miyasaka refined and discussed appropriate and inappropriate practices in these five areas. For example, teaching and/or practicing the entire domain of content objectives including the test objectives during regular classroom instruction could be appropriate if these are aspects of the curriculum. Whereas teaching and/or preparing for a test using highly similar or actual test content was considered inappropriate and unethical, as it could artificially raise test scores and decrease the validity of the test results (p. 6). Miyasaka's framework suggests that high-quality teaching of the content objectives and item formats during regular classroom instruction was test preparation of the best kind and minimised the need for extended test preparation.

To summarise, all of the frameworks presented above provide some fundamental insights into the understanding of the main elements of test preparation activities, thus serving as an important basis to draw upon for the coding of variables related to test preparation in the meta-analysis (Paper 1) of this research programme. For example, in the meta-analysis, the extent to which test preparation aligned with the test, and whether the intervention provided instruction on test-taking skills and/or socio-affective strategies (i.e.,

motivate students to take the test) were coded. These codes were consistent with categories described in the frameworks.

It should be noted that the three frameworks describing test preparation are mainly based on teachers' practices and perspectives, with students' participation in test preparation being overlooked. Some researchers have found it helpful to differentiate teacher-driven practices from student-driven practices (Anastasi, 1981; Arendasy et al., 2016; Briggs, 2009). Teacher-driven test preparation mainly includes coaching and curriculum alignment; while student-driven practice includes a variety of activities such as restudying materials, practicing sample items, using test-taking strategies, and asking for teachers' or peers' help (Briggs, 2009; Xie, 2013). In following sections of this chapter, the use and effect of teacher-driven practices and student-driven practices are reviewed separately.

2.1.3 Framework of test preparation in this research

Based on the review of terms uses in the literature and relevant frameworks on test preparation, a more comprehensive framework has been generated in this research as a synthesis of descriptions about test preparation in the literature, which describes the whole process of test preparation that includes actors, contexts, types, materials and outcomes of test preparation (see Figure 1). Actors of test preparation are mainly teachers and students, but other stakeholders such as parents and school administrators might also play roles in the test preparation processes, such as motivating students or monitoring teachers. Test preparation happens not only in formal school settings, but also in commercial institutes or at home. Three main types have been identified based on the understanding of how previous researchers identified this problem. School courses and commercial courses are two major types of test preparation that have been investigated widely in the literature. In schools, teachers might embed test preparation activities in the curriculum, or provide after-school courses for students who have the needs (e.g., remedial courses for left-behind students).

Commercial courses and private tutoring are another common type of test preparation. Students' self-preparation has received relatively scant research attention, though is very common as it happens both in schools and at home. During test preparation, a wide variety of strategies are used to study relevant materials, including but not limited to test preparation workbooks, online resources, practice tests and sample items, and so on. At the end of the test preparation journey, students take the test. The whole test preparation process produces an effect on students' test performance. In the whole test preparation process, every component interacts closely with others and contributes to the direction and magnitude of the test preparation effect on test performance. The effect of test preparation (i.e., outcomes) can also impact on the process in reverse. For example, unsatisfactory outcomes may in turn motivate teachers and students devote to a new round of test preparation. Therefore, a thorough understanding of the test preparation process is the prerequisite of a thorough understanding of test preparation effect, and vice versa.

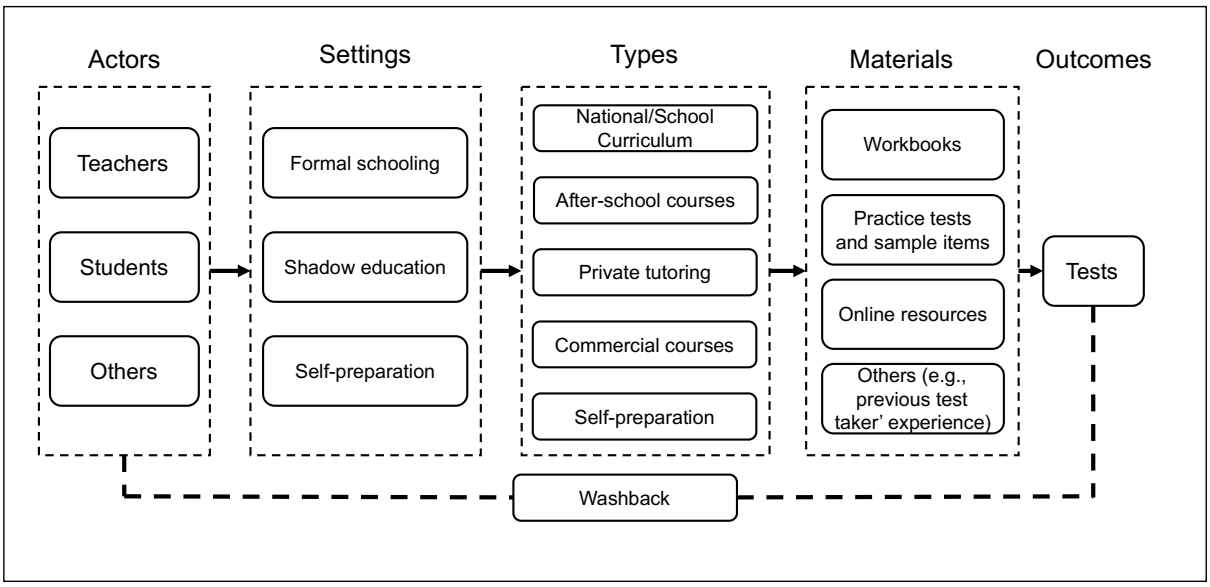


Figure 1 The whole process and components of test preparation.

2.2 Empirical studies on test preparation practices

Teachers and students engage in test preparation practices when they feel a need to achieve the best results in the test (Briggs, 2009; Crocker, 2003). This section reviews teachers' and students' test preparation practices identified in empirical studies.

2.2.1 Teachers' practices

Previous studies have suggested that in schools, teachers' instructional practices have been skewed toward test preparation when they face pressure to raise students' test scores (Au, 2007; Madaus, 1988; Nichols & Berliner, 2007; Popham, 1991). Many empirical studies found that to prepare their students for tests, teachers tended to narrow their teaching to tested content ('teaching to the test'), tailoring the curriculum to the test format and test content and providing a review of tested topics (Firestone et al., 2004; Shepard, 1990; Smith, 1991). Studies conducted in a wide range of countries, including Australia (Saif et al., 2021), China (Ma & Chong, 2022), U.S. (Firestone et al., 2004), Egypt (Gebril & Eid, 2017), Iran (Sadeghi & Ketabi, 2014), Germany (Jäger et al., 2012) etc., found that teachers generally taught students test-taking strategies before a test. Other test preparation activities were also commonly used by teachers, including teaching students time management skills, and providing students with test samples or test-like items to practise (Saif et al., 2021).

Researchers have found that teachers' decisions to offer test preparation activities in formal schools were influenced by the level of pressure they experienced to increase test scores (Firestone et al., 2004; Lipman, 2002; Smith, 1991). Such pressure may stem from the testing policies and can incentivise teachers to prioritise a narrow focus on the test over other forms of instruction. For example, Firestone et al. (2004) used a mix of surveys and direct classroom observations with interviews to examine teachers' test preparation practices in New Jersey. They claimed that teachers changed their teaching practices according to external state testing policies. When teachers felt higher pressure to increase test results from

school board administrators, they tended to use more conventional teaching with extensive drilling on test items. Firestone et al. referred to this type of teaching as 'didactic instructional strategies'. In contrast, when teachers received more external support and gained better understanding of state standards, they were more willing to provide more constructive instruction on curriculum subjects that allowed students to explore their ideas and build their own understanding of the knowledge and content. This type of teaching is referred to as 'inquiry-oriented instruction' (p. 67).

Teachers tended to tailor their instruction for high-stakes tests in response to the accountability pressures associated with testing, particularly in low achieving schools and before the test. Lipman (2002) found that teachers from low achieving schools focused more intensely on test preparation practices and were more likely to report employing them within their classrooms. Besides, in Gebril and Eid's (2017) study, teachers were found to increase the use of test preparation activities in their classes as the test date drew nearer.

As the literature suggests, formal school teachers' test preparation practices can be influenced not only by accountability pressures they experienced but also by the way test specifications were configured. Miyasaka (2000) suggested that when a test containing the same items is administered every year, most test information will be known to teachers and students. In this case, it is highly likely that teachers would choose 'teaching to the test' or even teach test items directly. By contrast, when the assessment objectives and the format of test items vary every year, teaching test content is almost impossible, thus there is more potential to promote a broader focus on curriculum objectives in teaching and test preparation. This is in keeping with Baird et al. 's (2014) study on the predictability of tests. In their report, Baird et al. examined the predictability of a leaving certificate test in Ireland and indicated that the scope of syllabus content and the alignment between syllabus and test papers determined the predictability of the test to a large extent, which further directed

teachers' practices in classrooms. It is possible that when a test is more predictable, teaching practices are more likely to be test-oriented.

Unlike teaching in the context of a school curriculum, teaching in special test preparation courses (either commercial courses or after-school courses provided by formal schools) or private tutoring have been found to be geared more specifically to improving test performance. Hayes and Read (2004) investigated activities in International English Language Testing System (IELTS) preparation classes by comparing two IELTS preparation courses in New Zealand – one independent course and the other as an additional module of a general English course. They found that teaching practices in those two courses were substantially different in terms of focus, delivery, and allocation. The independent IELTS preparation course was entirely focused on the test, that teachers provided activities exclusively around IELTS-like tasks, especially practice tests to help students prepare for the test. In comparison, the other teacher from the general English course introduced a wider range of activities in the class, including not only practice tests but also general reading or writing tasks and discussions about issues related to the test topics.

Teachers of special test preparation courses offer more test-specific teaching, which is a key selling point to attract students and parents. Yu et al (2017) examined the preparation of Test of English as a Foreign language Internet-based Test (TOEFL iBT) speaking test in China, and reported that teachers from test preparation schools frequently took the TOEFL test to record test items. These items and teachers' test-taking experiences were shared with students as part of the special test preparation course. Drawing on Miyasaka's (2000) framework on the relationship between curriculum and test content, teachers' practices observed in the context of preparing for the TOEFL exam reported in Yu et al. (2017), while inappropriate and educationally indefensible, might artificially raise students' scores and

undermine the validity of score interpretation. The effects of attending special test preparation courses (i.e., coaching) are discussed in later sections in this chapter.

2.2.2 Students' practices

Students' test preparation practices for admission tests have been examined by numerous studies (e.g., Briggs, 2001; Domingue & Briggs, 2009; O'Sullivan et al., 2021). Some included a large sample size and therefore reported potentially more robust findings. For example, Allensworth et al. (2008) examined practices of preparation for ACT in Chicago with 13,920 students participating in their survey. The results indicated that most test preparation practices occurred during class time, whereas more than half of the participating students reported preparing for ACT on their own by using test guidebooks or computer programmes. Students' preparation for ACT mainly involved doing sample test items and learning test-taking skills, such as skimming reading passages, moving ahead and process of elimination (Allensworth, et al., 2008).

Some studies from language testing also examined students' test preparation practices. For example, in their recent study about test takers' preferences for test preparation practices among prospective IELTS candidates from Asia and America, O'Sullivan et al. (2021) analysed questionnaires collected from 6,244 participants and found that students preferred a variety of test preparation practices, including face-to-face classes, videos and audios with tips, practice tests and online tutorials. No differences in students' preferences were found across geographical regions. Findings of Xie's (2013) research showed some variance regarding Chinese students' preferences of test preparation practices. Based on a questionnaire survey with 1,003 test takers of College English Test Band 4 (CET4) in China, Xie found that the most frequently used test preparation practice was rehearsing test-taking skills, followed by test preparation management, drilling, memorisation, and socio-affective strategies (e.g., seeking teachers' or peers' help). Strategies focused on the development of

language skills, such as reading extensively in English and regular English writing, were used scarcely.

Xie (2013) gave a further explanation of students' choices of test preparation practices. Phakiti (2003) differentiated language strategies according to their purposes and classified both use strategies, referring to strategies used to enhance performance, and learning strategies, referring to strategies used to enhance language learning and acquisition. Based on this categorisation, Xie argued that cognitive test preparation strategies such as drilling, test preparation management and memorisation are designed to enhance test performance, which could be regarded as use strategies. The development of language skills is used to enhance learning and acquisition, which could be regarded to as learning strategies. Xie found that test-takers may consider the differential functions of different strategies, and they chose use strategies more frequently in test preparation. This explanation is consistent with previous research findings on students' choices of strategies in learning.

In their research on qualitative differences in learning, Marton and Säaljö (1976) found that students adapted their learning strategies according to how they perceived test tasks and ways they would be assessed. This finding was later supported by a number of empirical studies. Based on questionnaires from 206 college students, Scouller (1998) found that students in higher education tended to employ surface learning strategies to prepare for a multiple-choice question examination, which they perceived as assessing lower levels of knowledge-based intellectual processes. As for the preparation for assignment essays, students were more likely to employ deep learning strategies as they believed that assignment essays required higher levels of cognitive processing. Broekkamp and Van Hout-Wolters (2007) summarised that students adapted strategy use according to demands of test preparation tasks, their perceptions of demands, personal goals, and ability to adapt strategies. In addition, Xie and Andrew (2013) found that perceived test value affects students' choices

of test preparation strategies. They administered a perception of assessment questionnaire at the beginning of a ten-week test preparation period and a test preparation questionnaire eight weeks later to 870 test-takers of CET4 in China. Results showed that test takers who perceived the test positively and had higher expectation of test success (higher self-efficacy) showed more engagement in test preparation (Xie & Andrew, 2013).

Moreover, a small number of researchers reported that students' use of test preparation was associated with their ability level and prior experiences. Gan (2009) surveyed students' IELTS preparation experiences in Hong Kong universities and found that low-achieving students were more likely to take test preparation courses prior to the test. Kitsantas (2002) found that high-performing students reported more self-regulatory strategies during test preparation, including goal setting and planning, self-monitoring, self-evaluating, self-consequencing, environmental structuring and help-seeking. During the test-taking processes, students who achieved high scores planned their responses to a larger extent (e.g., making an outline of the key points) and reviewed or revised answers frequently during testing.

To sum up, relevant studies provided explicit descriptions of teachers and students' test preparation practices. Empirical evidence has been collected in a variety of test contexts in different countries. However, existing studies mainly drew upon data from one specific test or a certain type of tests (such as language tests). The differences in test preparation practices in different test conditions have not been thoroughly compared and analysed in the current literature. To fill this research gap, the current research programme will provide a synthesis to investigate the variance of test preparation effects caused by different practices in different assessment contexts, aiming thereby to improve the general understanding of the test preparation process.

Another gap in the existing literature is that little attention has been given to students' self-study for test preparation. Many studies have examined teachers' practices of preparing

students for test-taking, while only a few focused on students' self-driven preparation practices. Given that students are the most important actor in test preparation, neglecting students' practices may lead to an inadequate and biased conceptualisation of test preparation, for example, considering students as passive receivers in the preparation courses or attributing test preparation effects to teachers' efforts.

Besides, despite the ample literature on students' use of learning strategies, only a limited number of studies investigated them in test preparation situations. Although the primary purpose (if not the sole) of test preparation—to maximise test scores—sets its nature different from our broad notions of learning, appropriate test preparation should be regarded as a learning activity. Thus, the use of cognitive and metacognitive learning strategies when preparing for a test may have great potential to trigger positive impact on both test performance and knowledge development. Therefore, the current research is also designed to provide more empirical evidence to explore students' use of learning strategies during test preparation and to understand the internal decision-making process of their test preparation practices.

This section reviews literature relating to test preparation practices of teachers and students. Building on this, in the next section, a review of the effectiveness of test preparation across different practices, individual characteristics and test contexts is presented.

2.3 Effects of test preparation

Studies on the effects of test preparation have proliferated since the 1980s, especially in the U.S. Extensive research has investigated if, and to what extent, test preparation might influence students' test results (e.g., Kulik et al., 1984; Montgomery & Lilly, 2012). Researchers have explored test preparation effects either through empirical research (e.g., Alderman & Powers, 1980; Briggs, 2009; Powers & Rock, 1999) or meta-syntheses and

meta-analyses (e.g., Bangert-Drowns et al., 1983; Becker, 1990; Montgomery & Lilly, 2012). However, existing research has not reached a consensus regarding test preparation effects. Some authors claimed that test preparation had a significant effect on test score gains (e.g., Briggs, 2001; Domingue & Briggs, 2009; Kulik, et al., 1984; Montgomery & Lilly, 2012); while others found no or extremely small effects (e.g., McGaghie et al., 2004; Powers & Rock, 1999). Several factors related to test preparation practices, contexts of tests, and individual differences have been identified by researchers as contributing to the differences in test preparation effect (Abrams et al., 2003; Appelrouth et al., 2017; Jäger et al., 2012; Montgomery & Lilly, 2012). The following sections review relevant literature on how these factors are related to varied test preparation effects.

2.3.1 Preparation practices

Effects of different test preparation practices

As mentioned earlier, special test preparation services, including both **commercial preparatory courses** and **private tutoring**, often assert their effectiveness on helping students maximise their test performance. For example, the Princeton Review, one of the largest test preparation companies in the U.S., offers a 'SAT 1400+' programme for those who have previously obtained a SAT score of at least 1250 on the Critical thinking and Math composite, which guarantee a score of 1400 or above (The Princeton Review, 2021). However, relevant research on test preparation effects generally reported small effect sizes (e.g., Briggs, 2001; Domingue & Briggs, 2009; Griffin et al., 2012, Zhang et al., 2022), and some studies even reported none or negative effects (e.g., Trenkic & Hu, 2021). For example, Domingue and Briggs (2009) used data from National Educational Longitudinal Study in the U.S. and found that enrolment in private tutoring and commercial coaching courses had small effects on SAT-Mathematics scores and no significant effect on SAT-Verbal scores. Similar results were reported in a study conducted in China. Based on data collected from one

Chinese city (N = 6043), Zhang (2013) found that private tutoring did not have a significant effect on the total score of the National College Entrance Examination (the *Gaokao*). Zhang suggested that a plausible reason was that the main function of private tutoring was to help students manage their time and provide more excises to get students familiar with the *Gaokao*. Practices provided by private tutoring might be similar to the self-study practices of students who did not attend private tutoring.

In general, studies suggest that test preparation courses offered by commercial providers might not have the extraordinary effect they claimed to have. Regarding test preparation practices in the context of formal schooling, most studies investigated the effect of a special test preparation course and reported significant effects, though the effect sizes varied (e.g., Alderman & Powers, 1980; Filizola, 2008; Gaye, 2001). In addition, some studies examined the effect of specific practice in schools, such as providing workshops on **test-taking strategies** (e.g., Samson, 1985; Takallou et al.,2015) or providing students with **practice tests** (e.g., Bol & Hacker, 2001). For example, Takallou et al. (2015) used an experimental study to examine the effect of teaching test taking strategy on Iranian high school students' performance on English section of the University Entrance Examination (Konkoor). They found that students who participated in the test-taking strategy workshop (the treatment group) significantly outperformed the control group in the test.

In terms of students' use of **practice tests** in test preparation, its positive effect on test performance has been corroborated by previous studies. This effect is referred to as the 'testing effect' (see Hausknecht et al., 2007; Scharfen et al., 2018) or 'practice effect' (Kulik et al.,1984). The testing effect on memory retention has been widely observed in laboratory settings and in natural learning contexts (for reviews, see Karpicke, 2017; Scharfen et al., 2018). In authentic educational settings, the literature on the testing effect reported inconsistent findings (Bangert-Drowns et al., 1991; Moreira et al., 2019; Phelps, 2012). Some

studies reported significant effects of using practice tests on students' test performance. For example, Batsell et al. (2017) conducted a quasi-experimental study, in which they assigned two classes of students enrolled in an introductory psychology course to either a quiz group or a control group. Students in both groups read textbook chapters that were relevant to the content assessed by tests, and the quiz group also completed multiple-choice quizzes related to their reading. Students' performances on three course exams during the academic term were compared. The results showed that the quiz class scored significantly higher than the control class, and the differences were significant, both when the test questions were identical with the quizzed questions and when they were new questions from the textbook.

In contrast, there is increasing evidence that testing practices are not as effective or even inferior to restudy practices for tasks that require complex procedures, e.g., problem-solving tasks in mathematics or science (van Gog et al., 2015; van Gog & Kester, 2012; Yeo & Fazio, 2019). For instance, in a classroom-based study in the U.S., Bol and Hacker (2001) compared the effect of two different test preparation methods—practice tests and traditional review (e.g., lectures provided by teachers)—on the mid-term and final tests of a research methods course. They found that students in the review group only outperformed the practice tests group on multiple-choice items on the midterm exam, while there was no difference between the two methods on final tests. Therefore, the question of whether testing practice can retain its strengths for complex learning tasks that require not only memory retention but also transfer remains unanswered, and this research programme will provide some insights into this question.

Previous studies suggested that testing and restudy are two common strategies students used for test preparation (e.g., Karpicke et al., 2009). As mentioned above, the effect of testing has been widely explored in both laboratory and classroom settings. However, most experimental studies examined the effect of a certain strategy over single sessions. In

authentic educational contexts, it is unlikely that students would engage in one single acquisition session solely using one strategy during their test preparation process, as manipulated by experimental studies. To bridge this gap, some studies investigated the effect of **multiple uses of one strategy** or **combined use of multiple strategies** (e.g., McDaniel et al., 2015; van Gog et al., 2011). For instance, Gurung and Burns (2019) found that repeated testing led to lower final test scores. Similarly, McDaniel et al. (2015) found that compared to repeated testing, a combination of testing-restudying-testing can produce the greatest effects on students' learning. In addition, Van Gog et al. (2011) suggested that using different combinations of learning strategies in certain sequences may result in different outcomes. They examined students' performance on problem-solving tasks and found that students would benefit more from restudying-testing (i.e., studying example first, then solving problems) rather than testing-restudying (i.e., solving problem first, then studying examples).

The studies mentioned above suggested that using a combination of testing and restudy strategies in a certain sequence during test preparation may lead to better effects. However, since the number of studies examining this problem is limited, further research is needed to gain a more comprehensive understanding on how different combinations of strategies cause different effects on different tasks, and what factors contribute to the variability in these effects. Moreover, it is still not clear how students would use different strategies when they do self-study for a test without any manipulation, and what factors affect students' blended uses of the two strategies. To bridge this gap, this research programme examines students' combined use of testing and restudy strategies in test preparation contexts. Factors influenced students' strategy choices and their perceptions on the two strategies are investigated, as well as the effects of different strategy combinations. This research can be a valuable addition to the literature by extending the investigation of students' strategy use to test preparation context and taking students' perceptions into consideration.

The time factor of test preparation practices

Across different practices, an important feature that may affect their effects, as reported by some previous studies, is time. Existing research revealed that the effect size of test preparation is related to the **length of the programme** (e.g., Bangert-Drowns et al.,1983; Devine-Eller, 2012; Xie, 2013). It was believed that the length of test preparation intervention influences the effect size (Appelrouth et al., 2017; Ireson, 2004). Some researchers asserted that longer programmes led to more score gains (Bangert-Drowns et al., 1983; Samson, 1985). In a recent study on the effect of test of preparation on SAT test performance, Appelrouth et al. (2017) found that individual tutoring hours had the most significant effect on test score increase and so did the number of hours of group tutoring, although the latter effect size was smaller. They estimated that each additional hour of individual tutorial increased the final scores by 2.34 points on average, while the increase per additional hour for group tutoring was 1.48 points on average. However, when the test preparation takes place could also be important: Devine-Eller (2012, p. 463) reminded us that,

> *the preparation must be undertaken with the proper amount of lead time—not so little that it is rushed, and not so much that the skills training atrophies.*

That is to say, **the starting time for test preparation** might be a predictor for the score increase. Starting the preparation at an appropriate early time can yield a significant score increase (Appelrouth et al., 2017; Devine-Eller, 2012; Samson, 1985).

Additionally, **the distribution of test preparation time** has been observed as having effects on score gains (Appelrouth et al., 2017). Compared with students who received intensive test preparation instruction prior to the test, those who received distributed tutoring hours over time were more likely to make more score gains. Some researchers stated that this beneficial effect could be attributed to the working mechanism of long-term memory, as

complex cognitive skills and long-term retention knowledge require sufficient intervals to be encoded and transferred into long-term memory (Kapler et al., 2012). In this case, frequent input and stimulation might even hinder the coding process and obstruct the recall and application of knowledge.

To recap, the literature on the effects of test preparation practices can be divided into two main aspects. One part of the literature has investigated the effectiveness of teacher-led test preparation courses mainly by comparing the performance of students who attended the course with those who did not. Some studies within this area have explored how specific characteristics of test preparation courses, such as the length of the course, or the instruction on test-taking skills, can impact the magnitude of test preparation effects. A significant proportion of relevant studies has focused on this aspect, often based on large-scale or longitudinal datasets, or derived from experimental designs. The literature illustrates the need for an up-to-date synthetical study which integrates the effects reported by individual studies and comprehensively compares the effects of different test preparation practices (e.g., commercial courses or courses provided in schools). Findings of the synthesis may provide valuable information for determining the extent to which investments into commercial courses is justifiable.

In contrast, another set of studies has focused on students' uses of strategies, such as testing and combined use of testing and restudy, and their effects on test performance. This line of research has adopted a more nuanced approach, usually experimental designs, to investigate the effect of strategies at a granular level. However, it is important to note that most studies in this area have tended to contribute to our understanding of how the adoption of specific strategies facilitated students' learning or abilities (e.g., problem-solving ability) in a broader sense, while only a limited number of these studies has been conducted specifically in an authentic test preparation context (e.g., Janes et al., 2018). As noted earlier,

test preparation is a unique setting of learning, where students' motivation, beliefs, and study behaviours might be different from ones observed during more typical learning settings. In some contexts, the ultimate goal of test preparation is to improve test performance which will allow students to move to the next stage (e.g., pursue their studies in higher education). Therefore, the evidence reported in current literature lacks ecological validity because most of it was not conducted in authentic test preparation settings, using real-life educational assessment materials. It is essential to identify and evaluate effective strategies specifically tailored to the purpose of improving test performance in this context, as they might not be equally effective in both test preparation settings and more typical learning settings.

2.3.2 Different contexts of tests

Research findings indicated that the overall test preparation effects were inconsistent across different tests. A large number of studies examined the test preparation effect for college admission tests, especially in the U.S. context (e.g., Alderman & Powers, 1980; Briggs, 2001; Dale, 2013; Holmes & Keffer, 1995). Bangert- Drowns, et al. (1983) conducted a meta-analysis of the effects of 30 controlled studies on test preparation programmes for achievement tests, finding that test preparation had a positive effect on test scores of approximately a quarter of standard deviation (*SD*). However, Scholes and Lain (1997) asserted that test preparation had only a minimal impact on ACT score gains beyond gains from simply retaking the test. In terms of aptitude tests, Kulik et al. (1984) reported that the average effect of test preparation was 0.15 *SD*s on SAT and 0.43 *SD*s on other aptitude tests.

There were also some studies looking at test preparation effects in other tests. For example, McGaghie et al. (2004) reported an extremely small and non-significant effect of commercial courses on medical examinations, based on the data extracted from 11 empirical studies. As for language tests, a recent study suggested a weak relationship between test preparation and TOEFL performance (Yu et al., 2017) yet earlier studies reported varied

effects of test preparation. For instance, Elder and O'Loughlin (2003) reported that students gained about half a band score in IELTS after attending a nine-month test preparation programme. However, Nguyens's (2007) investigation of the effect of IELTS and TOEFL preparation courses demonstrated that test preparation did not affect test-takers' listening test scores. Green (2007) investigated the effect of three types of English courses on students' performance of IELTS writing – a dedicated IELTS preparation course, a pre-sessional course in English for Academic Purposes (EAP), and a course that combined features of both. He found that all three types of courses improved students' test scores, while the dedicated test preparation courses did not improve writing test scores to any greater extent than others.

It is worth noting that studies mentioned above were all conducted in high-stakes testing contexts, where test results have important consequences for students. Only a few studies in the field examined test preparation effects in low-stakes testing contexts, where the results of the test have no personal relevance for test-takers. For example, Brunner et al. (2007) conducted a study in two secondary schools in Germany to examine the coaching effect for the PISA (Programme for International Student Assessment) test, which is a low-stakes international test assessing 15-year-old students' academic achievements in reading, mathematics, and science. Brunner et al. guided teachers to dedicate four PISA-related courses (each lasting 45 minutes) on reading and mathematics. They assessed the impact of coaching on students' performance using a pre-test and post-test design. Both the pre- and post-tests were composed of items from PISA 2000 item pool. Researchers found that coaching had small to medium effects on mathematics (d = 0.16) and reading (d = 0.43) for students in high academic tracks, while it had no effect (d = -0.03 and 0.07 for mathematics and reading, respectively) for students in low academic tracks.

In a more recent study, Dahlke et al. (2023) used longitudinal data (N = 120,384) to compare students' improvement in low-stakes and high-stakes tests in the U.S. In this study,

students' score gains from 9th grade to 10th grade on Preliminary Scholastic Assessment Test (PSAT) were regarded as the reflection of improvements in low-stakes settings, while score gains from 11th grade to 12th grade on SAT reflected score gains in high-stakes settings. The results showed that the levels of score improvement in the two settings were similar.

In summary, variability in the effects of test preparation on test score gains across different types of tests reveals that some tests might be more amenable to preparation than others. Hence, conducting a meta-analysis that compares test preparation effects on various tests can provide a clear understanding of the moderating effects of test types. Besides, it remains unclear whether the test preparation effect differs for tests with different levels of stakes, as findings in studies are inconsistent. To fill this research gap, the meta-analysis and experimental studies in the research provide more evidence on the impacts of the levels of stakes of tests on the effect of test preparation. Factors related to tests are not the only moderators of test preparation effects, research suggested that individual characteristics such as students' ability and experience can also influence the extent to which test preparation affects score gains. Factors of individual differences are reviewed in the next subsection.

2.3.3 Individual differences

Researchers have identified several personal factors that are related to test preparation effects, such as test- taking experience (e.g., Trenkic & Hu, 2021), prior academic attainment (e.g., Griffin et al., 2013), and students' perceptions of tests (e.g., Hong & Peng, 2008; Xie & Andrews, 2013). Some of these factors can be associated with the test preparation practices that students use, how students engage in preparation, and the outcomes that they may achieve as a result. For example, students' motivation for test preparation and test taking might have an indirect or mediating effect on test performance, as studies have suggested that motivated test-takers outperformed unmotivated students (Baumert & Demmrich, 2001; Cole et al., 2008; Penk & Richter, 2017). Hong and Peng (2008) examined the extent to which

Chinese 7th and 11th graders attached value to test and the effect on test performance. They found that students' perceptions of tests had a significant direct effect on motivational and metacognitive regulation in students' test preparation as well as an indirect effect on test performance. Similarly, as mentioned previously, Xie and Andrew (2013) found that Chinese university students who endorsed positive value of test and had high success expectations tended to be engaged more actively in test preparation.

In addition to the value that students attached to a test, students' prior attainment and test-taking experience may also be associated with test preparation practices and test performance. Sesnowitz, Bernhardt, and Knain (1982) analysed data of 2,240 uncoached and 513 coached students in the New York metropolitan area and found that students with previously poor test performance had significant increases in SAT scores by coaching. Griffin et al. (2012) investigated whether attending commercial test preparation course affects performance in the Undergraduate Medical and Health Sciences Admission Test (UMAT). They found that enrolment in commercial coaching courses had positive effects for high ability students on their performance in non-verbal or abstract reasoning parts in the test and negative effects for lower ability students.

Moreover, a limited number of qualitative studies suggested that students' adoption of different test preparation and test-taking strategies may also lead to different effects. Damankesh and Babaii (2015) adopted a think-aloud approach to track Iranian students' test preparation and test-taking strategies during their high school final language test. They found that students who received higher scores preferred using strategies such as lexical cohesion, eliminating bad alternatives, guessing intelligently, employing reasoning and activating knowledge of collocations. By contrast, low achievers admitted that they guessed randomly, copied words and phrases from previous items or alternatives, responded by chance or even cheated when they were taking the test. Consequentially, researchers inferred that students

with higher academic achievement may benefit more with test-taking strategies than their low achievement peers, as they showed more interest in using effective strategies with higher-cognitive demand and utilizing cognitive monitoring to increase the possibility of success (Damankesh & Babaii, 2015; Kitsantas, 2002). To note that this inference of students' prior attainments and their preferences for strategies was based on qualitative research, while the field lacks sufficient quantitative research to verify the inference. To address the research gap, this research programme provides some quantitative evidence on the differential utilisation of test preparation strategies by students with varying prior achievement levels, as well as the potential relationship between variations in strategy uses and score gains.

This section reviewed the effects of test preparation with regard to different practices, contexts of tests, and individual differences. To explain the observed effects, several theories have been proposed and employed in previous studies. The following section highlights the theories that will be used throughout this thesis to interpret the findings of the studies carried out in this research programme.

2.4 Theoretical underpinnings for the test preparation effect

In this section, theoretical underpinnings that can help explain the mechanism of test preparation are reviewed. The framework proposed by Arendasy et al. (2016) on causes of test preparation effects is presented first. Then, various models in the field of cognitive and metacognitive science are visited, including cognitive load theory, transfer-appropriate framework, knowledge-learning-instruction framework, and theories on metacognition, to explore whether test preparation effects can be explained through these lenses.

2.4.1 Causes of test preparation: Arendasy et al. (2016) framework

Arendasy et al. (2016) summarised four possible models from the existing literature (e.g., Anastasi, 1981; Messick, 1982; Lievens et al., 2007) to explain the causes of test preparation

effects. The first postulated that score gains should be attributed to **familiarisation with test-taking and test structure**, which reduces construct-irrelevant variance. Researchers hypothesised that students who are not familiar with standardised testing need to familiarise themselves with the test when solving the first few test items. As a result, the first few test items are harder for them than for those students who have certain levels of familiarity with the test. In this case, test preparation can reduce the detrimental effect of individual difference in test familiarity, but such score gains may not reveal an increase in students' abilities, because they do not result from an increase in knowledge or skills that are relevant to the construct being assessed.

The second model states that engaging in test preparation improves **test-wiseness and test-taking skills**, such as guessing and eliminating wrong answers. Arendasy et al. claimed that these tactics can improve the likelihood of correctly solving test items, without affecting the underlying latent traits.

The third model ascribes score gains to teaching **construct-relevant materials**, which further results in the development of test-specific cognitive abilities or an increase in domain-specific knowledge.

The fourth model is based upon the premise that the increase in test scores is caused by the improvement of more **generalisable cognitive abilities or knowledge structures**, such as higher-order knowledge factors.

These four models form various assumptions of the effect of test preparation. However, Arendasy et al.'s (2016) summary was mainly drawn from retest effect and cognitive ability testing literature, its applicability in real test preparation contexts needs further investigation.

As mentioned earlier, test preparation can be considered, among many other things, a learning process albeit with some special features, such as greater intensity and a narrow focus. Therefore, learning theories in the field of cognitive and metacognitive psychology can

be useful to interpret test preparation effects, as well as to provide guidelines about how to improve the effectiveness of test preparation activities.

2.4.2 Cognitive Load theory (CLT)

A helpful cognitive learning theory that can provide insights to the effectiveness of test preparation practices is CLT. It is a well-established framework that explains the cognitive processes involved in information processing and their impact on learning (Paas et al., 2003; Sweller, 1994, 2010). CLT proposes that there is a limited number of cognitive resources available for processing information. The effectiveness of students' learning is influenced by how these resources are used. Three main types of cognitive load are posited, as shown in Figure 2: intrinsic, extraneous, and germane. **Intrinsic load** is the inherent difficulty and complexity of the material to be learned, **extraneous load** refers to the manner in which the material is presented, and the **germane load** refers to the amount of effort required to make connections and build new schemas, which is considered to be a positive type of cognitive load and can result in the formation of durable and meaningful connections in students' memory.

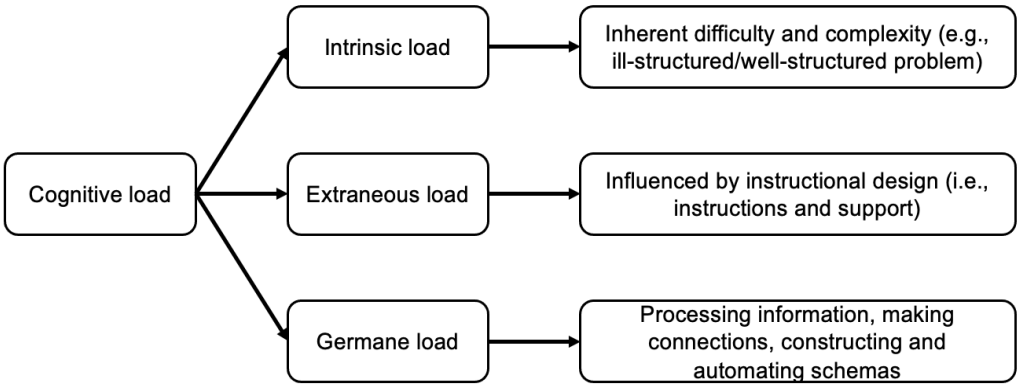


Figure 2 Structure of Cognitive Loads 2.

CLT suggests that it is important to reduce extraneous load and increase germane load to improve the effectiveness of learning (Paas et al., 2003, 2004). In the context of test

preparation, this can be achieved by designing instructional materials that are clear, concise, and visually appealing and by using sample items and working examples that help students understand the materials and make connections to tasks in the test. Besides, students' experiences and prior knowledge as well as the nature of the to-be-learned materials should be considered when designing test preparation instructions, to optimise the allocation of cognitive resources and maximise the learning outcomes.

CLT can be used to understand the effects of test preparation by considering different types of cognitive load and cognitive resources that are involved in the process of preparing for a test and taking the test. However, as discussed above, the alignment between test preparation and the tests can have an impact on test results: test preparation with a narrow focus (e.g., specific coaching courses designed for the test) may lead to better test performance. While CLT mainly focuses on managing cognitive load during learning, it does not directly inform how to explain the effect of different alignment levels between test preparation and tests. The Transfer-Appropriate Processing framework can provide some insights in this regard.

2.4.3 Transfer-Appropriate Processing (TAP) framework

The TAP framework has been widely used to explain the advantages of testing practice on memory recall (e.g., Yang et al., 2021). TAP mainly states that the memory recall is affected by the degree to which the processes engaged during encoding match the processes engaged during retrieval (Blaxton, 1989; Morris et al., 1977). In other words, if the way information was processed during learning is appropriate for the way it will be used in the test task, better retention is more likely to occur. According to TAP, the effect of test preparation can be boosted if processing strategies used during test preparation match with those used during tests—using testing practices during the preparation stage will yield more benefits (Veltre et al., 2015). As discussed in earlier sections, the effects of testing practices have been widely

confirmed by a substantial number of studies conducted in both laboratory and classroom settings in recent years (e.g., Carpenter et al., 2009; Goossens, N., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H., & Zwaan, R, n.d.; Rawson & Dunlosky, 2012; Roediger & Karpicke, 2006; Rowland, 2014). However, it is still not clear whether the TAP framework can be used in broader learning contexts beyond memory recall. Further research is needed to determine the potential applications and limitations of TAP in complex learning situations.

2.4.4 Knowledge-learning-instruction (KLI) framework

A more recent framework, the KLI framework proposed by Koedinger et al. (2012), shares some similarities with the TAP framework in their emphasis on the importance of alignment between different aspects of learning process and assessment process. KLI highlights the fact that cognitive science and education research have traditionally focused on the most observed educational events, including *instructional events* (e.g., explanation, lecture, text) and *assessment events* (e.g., test, discussion), which also could be easily manipulated by experiments. The unobservable underlying *learning processes* (e.g., memory and fluency building, induction, sense-making) and *knowledge components* (e.g., facts, schemas, procedures, principles) have been left out.

According to the framework, there are different types of *knowledge components*. A common type of knowledge has constant application conditions or cues and constant target responses, which is known as **constant-constant knowledge components**. For example, learning that the Chinese radical '日' (constant cue) means 'sun' in English (constant response), or learning that the 'circle area formula' (constant cue) is $A = \pi r^2$ (constant response). Constant-constant knowledge components require predominantly memory processes.

In contrast, knowledge that may have variable application conditions and variable responses is known as **variable-variable knowledge components**. For example, learning

that there are various spellings for plural English (variable conditions) and the plural of a word that ends in a consonant plus ‘y’ can be formed by changing ‘y’ to ‘ies’, and simply add an ‘s’ to the end of the word if the word ends in a vowel plus ‘y’ (variable responses). For variable-variable knowledge components, their acquisition predominantly requires induction (i.e., drawing general principles or rules from specific examples) or compilation (combining smaller pieces of knowledge into a larger or complex structure) processes. There are also **constant-variable knowledge components** involving recognising or classifying different types of objects or stimuli, and **variable-constant knowledge components** which involve applying a set of rules or procedures to solve a problem.

This framework suggests that the fit between the nature of knowledge components and the learning processes required to acquire the knowledge is critical for learning outcomes. When it comes to test preparation, it is important to consider what type of knowledge components are being assessed, as test preparation strategies may have different effects on different types of knowledge components. For instance, if a test primarily measures constant-constant knowledge components, then test preparation practices that focuses on memorisation may be effective. This should be taken into consideration when teachers design courses and when students plan their studies when preparing for tests.

2.5 Summary

In general, there is a rich set of test preparation studies that has been conducted across different countries and different educational levels. Existing studies have provided adequate and valuable evidence on the effects of different practices employed to prepare for different kinds of tests. Yet there is still muddiness in the literature regarding the mechanisms of how test preparation works, as well as under which conditions and which test preparation practices are more advantageous. Individual empirical studies have provided very limited insights on

these questions, which is understandable to a large extent, as it will be too demanding for an individual study to take comprehensive and systematic considerations of all possible factors (e.g., features of test preparation activities, participants' characteristics, material use, instructions, and learning strategy use, etc.) that might alter the effect of test preparation interventions. It is highly possible that many studies disregard these areas due to the difficulty of separating single factors from the integral process and examining their contributions separately. In this case, it is essential to understand every component of test preparation processes (see Figure 1) and to clarify how these components influence the effects of test preparation individually or jointly. Therefore, a quantitative systematic synthesis is required, as it not only provides a comprehensive understanding of an overall test preparation effect on educational tests, by the extracting information across different studies, but it also depicts a clearer picture of the extent to which different components in the test preparation process contribute to the overall effect. As most existing synthetical reviews were published decades ago and focused on the effect of coaching (e.g., Bangert-Drowns et al., 1983), the meta-analysis in the current research programme can provide more up-to-date and holistic conclusions on test preparation effects.

Another research gap identified by this chapter is that most studies examined the effects of teachers' practices rather than students' practices, and the effects of large-scale tests rather than small-scale tests. However, in many cases, students' test preparation has been unsupervised, especially in contexts of higher education, small-scale or low-stakes tests. In the field of cognition and learning, researchers have investigated the effectiveness of learning strategies used by students. Testing and restudy have been two strategies that have been reported to be commonly used by students for test preparation. The effectiveness of these two strategies has been initially explored in the field of memory and retention in a large number of laboratory studies. Later, a similar design has been applied in classroom contexts. Some

studies confirmed the beneficial effects of testing practices on students' test performance and long-term learning. However, most classroom-based studies used highly similar or identical materials in the intervention and measures, which results reflected the effect of the two strategies on knowledge recall. The (quasi-) experiments in the research programme examine the transferability of the effects of testing and restudy on educational test performance, using real-life test preparation and assessment materials that focused not only on knowledge recall but also knowledge understanding and application, thus can provide important and practical insights into the effectiveness of students' use of the two strategies in real test preparation contexts. Additionally, the last experiment in this research depicts a complete portrait of how students made decisions on the use of the two strategies and provides qualitative data to help interpret students' strategy decisions. This experiment is also a crucial addition to the inadequate literature on students' self-driven test preparation practices and their perspectives on this process.

This chapter has provided an overall review of empirical and theoretical literature related to test preparation effects. In essence, this chapter synthesises the key components of the test preparation process and discusses how these components contributed to its effectiveness. Therefore, the literature reviewed in this chapter mainly includes work that can contribute to the overall understanding of the test preparation process and effects. Three types of literature have been left out from this chapter to avoid repetition: 1) experimental studies that quantified the effects of different test preparation practices in different context of testing; 2) studies that compared the effectiveness of testing practices and other learning activities in both laboratory and classroom settings; and 3) literature on the role of metacognition during test preparation process. The three types of literature are discussed in papers presented in Part II of the thesis, because they inform the design and contribute to the in-depth discussions of the specific research questions in those papers.

Chapter 3. Methodological Strategy

This chapter begins by outlining the overall design of the research which consists of a meta-analysis and four interrelated (quasi-) experimental studies. The rationale of each study is clarified and the underlying connections between studies are explained. To avoid redundancy, this chapter does not delve into the detailed methods of each study since these are presented in each paper in the second part of the thesis. The overall methodology of the four (quasi-) experiments conducted in the same research site is comprehensively articulated in this chapter. At the end of this chapter, ethical considerations are discussed.

3.1 Rationale of the research

This research programme includes a meta-analysis of the test preparation effects observed in large-scale educational tests, as well as four related (quasi-) experimental studies that examine the effects of students' strategy use, students' study behaviours and their perceptions on test preparation strategies.

The literature review in Chapter 2 revealed variations in the magnitude of the test preparation effects across different practices and contexts, and a lack of comprehensive investigation into the mechanisms underlying the effect. Moreover, previous reviews focused mainly on coaching effects in the U.S. context, and most were published several decades ago, highlighting the need of an up-to-date and comprehensive research synthesis on the topic. To address this gap, a new, comprehensive meta-analysis of test preparation effects on large-scale educational tests has been firstly conducted in this research. This meta-analysis provided an average estimate of the effect of test preparation as well as an overall understanding of relevant moderators associated with the direction and magnitude of the effect. Besides, the preliminary results of the meta-analysis revealed some gaps in the literature and have guided the design of subsequent (quasi-) experiments.

To address gaps identified through conducting the meta-analysis and literature review in Chapter 2, the subsequent (quasi-) experiments designed more ecologically valid interventions where students led their self-study for test preparation using different strategies. These (quasi-) experimental studies were conducted in China, an educational context and culture that is very different from the West. Thus, studies in this research can make a significant contribution to the literature by adding empirical evidence from Chinese contexts to the predominantly Western (more specifically the U.S.)-led field. The effects of these student-led, self-paced practices are examined by the (quasi-) experiments, as well as students' individual differences and their perceptions of their test preparation practices. Moreover, by exploring the effect of learning strategies (i.e., testing and restudy) used by students on their test performance, the four (quasi-) experiments also contribute to the field of learning strategies by expanding the investigation of their effectiveness to test preparation contexts.

In total, through integrating synthetical evidence from the meta-analysis and empirical evidence from (quasi-) experiments, this research can provide a more comprehensive understanding of the effects of test preparation and students' strategy use for test preparation.

3.1.1 Rationale of the meta-analysis

As discussed in Chapter 2, the effect of test preparation has been investigated through a considerable number of empirical studies, but findings reported in these studies were inconsistent. Note that in the relevant literature, only a few studies used an experimental design; most studies used self-report surveys to investigate participants' test preparation histories, then compared the test performance of students who engaged in test preparation to that of students who did not prepare for the test (e.g., Briggs & Domingue, 2009; Powers & Rock, 1999; Hu & Trenkic, 2019). Specific characteristics of test preparation activities, such as material use, duration, and students' strategy use, etc., were rarely collected and reported

in most self-report surveys. Thus, the effect of test preparation presented by these studies was possibly inaccurate, as the confounding variables were not well controlled, and the heterogeneity of test preparation practice may have been overlooked. Missing in the literature is a research synthesis that provides a comprehensive picture of test preparation activities and their effects.

To fill this gap, a new meta-analysis has been conducted at the beginning of the research programme. The meta-analysis systematically synthesises results of several independent, combinable studies, leading to a quantitative summary of the pooled results (Egger & Smith, 1997). This statistical analysis allows for the interpretation of the effectiveness of an intervention. This task can be difficult to achieve using other methods such as narratively summarising findings, especially when the number of studies is high.

Furthermore, a meta-analysis also increases the precision and generalisability of the estimated effect size, as it enables the exploration of which factors predict the variability in the magnitude and direction of the intervention effect, thus generating a hypothesis about possible moderators and directing future studies (McKenzie, Beller & Forbes, 2016). In this regard, conducting a meta-analysis first allows the researcher to build a comprehensive, transparent, and replicable framework of the test preparation effect, which cannot only serve as a foundation for subsequent studies but also support future evidence-based practice.

After a systematic literature search and screening, the scope of the meta-analysis was narrowed to examine the effect of test preparation on student performance on large-scale educational tests. The focus on large-scale assessments has been largely due to the vast amount of information and research available around them. Indeed, the majority of relevant studies identified in the systematic literature search focused on large-scale tests, particularly college admission tests and English proficiency tests. The large amount of data available makes it possible to gain substantial data to carry out the meta-analysis.

Additionally, information and resources related to large-scale educational tests are widely available to the public. Thus, experienced teachers are familiar with tested content and test forms, and they can provide high-quality test preparation activities. Students can also utilise various test preparation materials such as handbooks and sample tests to direct their own preparation for the test. In this case, extracting features from these activities and investigating how different characteristics of test preparation activities contribute to the preparation for a specific test would be valuable. The focus on large-scale educational tests of the meta-analysis can produce precise results, which can be generalised to the wider population of test-takers of large-scale educational tests and to various educational contexts.

In total, 30 studies met all inclusion criteria and were included in the meta-analysis, reporting 97 effect sizes. As most included studies reported more than one effect size, and these effect sizes were estimated based on the same sample, the sampling errors were assumed to be correlated. Therefore, a three-level correlated and hierarchical effects model with robust variance estimation was adopted to calculate the overall effect size of test preparation interventions.

Moderating effects of possible moderators were examined by sub-group analysis and meta-regression to investigate if the variance of magnitude and direction of the effect could be explained by test and test preparation-related factors as well as other characteristics. Based on an iterative approach during the coding process, four types of moderators were identified and investigated by the meta-analysis, including test conditions (i.e., nature, domain, education level), test preparation characteristics (i.e., manner, material use, strategy use), student characteristics (i.e., ability level, previous experience), and design characteristics (i.e., allocation method, comparison conditions).

3.1.2 Rationale of the (quasi-) experiments

The meta-analysis identified various gaps in the literature, and hence informed the design of the four subsequent (quasi-) experimental studies in the research programme. One of the gaps was the scarcity of studies carried out in real contexts of test preparation. The four (quasi-) experiments addressed this gap by being conducted in ecologically valid settings. Another gap was the lack of studies examining the effects of specific test preparation strategies during test preparation. This research explores (quasi-) experimentally the effects of single and combined use of two test preparation strategies during students' self-study for the exam. A third gap identified by the meta-analysis is the scarcity of relevant studies carried out outside of the West. The (quasi-) experiments in this research were carried out in a Chinese assessment context and one quasi-experiment specifically examined how Chinese students regulated their use of the two strategies and their perceptions of the strategy usage during test preparation.

To preview briefly, the list of the four (quasi-) experiments are provided here, and an overview of the four studies are presented in Figure 3:

Experiment 1: The effects of using testing and restudy in a high-stakes context.

Experiment 2: The effects of using testing and restudy in a low-stakes context

Experiment 3: Metacognitive monitoring and control of test preparation

Experiment 4: Students' blended use of testing and restudying strategies

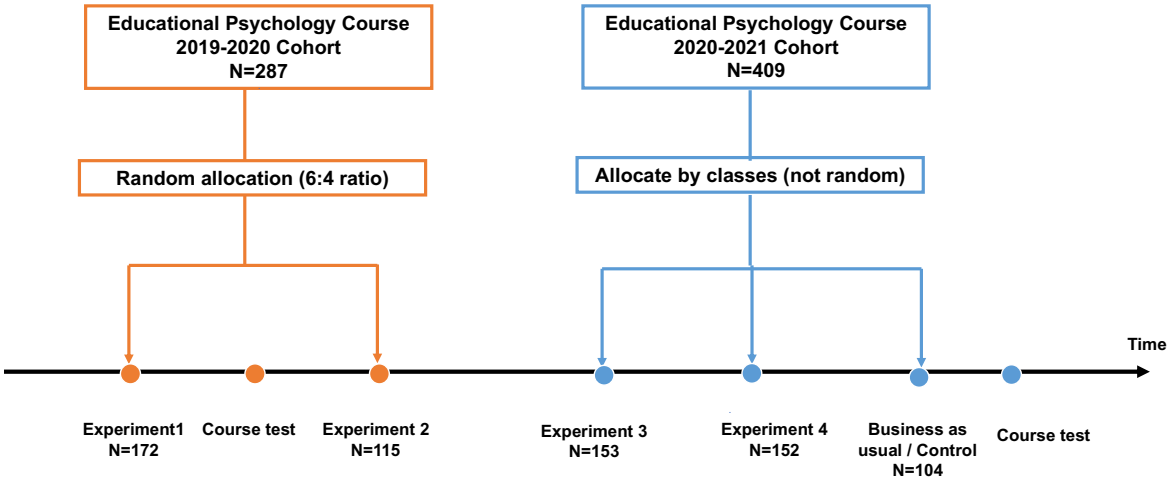


Figure 3 An overview of four (quasi-) experiments in the research.

The adoption of a (quasi-) experimental design was due to the nature of the gaps identified in the meta-analysis and the inability to draw conclusions about potential causal relationships that may exist between various test preparation practices, conditions and effects. Experimental studies have two fundamental features (Price et al., 2017). The first is that the researchers manipulate, or systematically vary, the level of the independent variables and thus create different conditions. The second is that the researchers control or minimize the variability in variables other than the independent and dependent variables, which are called extraneous variables. Lacking control on extraneous variables will make it difficult to separate the effect of the independent variable from the effects of the extraneous variables. In this research, (quasi-) experimental studies were carried out to specifically answer the questions of whether there is a causal relationship between two types of variables: the independent variables, i.e., test preparation strategies and conditions in this research, and the dependent variables, i.e., the effect of test preparation. The possible impact of other extraneous variables, such as students' prior attainment, was controlled. In some cases, random assignment is not possible. Thus, quasi-experimental designs are commonly employed in educational research (L. Cohen et al., 2007; Gribbons & Herman, 1997).

Although groups are not randomly assigned in quasi-experimental designs, researchers can try to account for any extraneous variables by controlling for them in the analysis or by choosing naturally occurring groups that are matched on as many variables as possible (e.g., gender, socioeconomic status, age, prior attainment, etc.).

In this research, Experiments 1 and 2 were true experiments using pretest-posttest control and experimental group design. The design of the two experiments is shown in Figure 4. The two experiments examined the effect of using testing and restudying strategies for test preparation in two different contexts: the high-stakes testing setting and the low-stakes testing setting. Participants in the two experiments were randomly assigned to one of the four experimental and control condition: *Testing only*, *Testing with feedback*, *Restudy,* and *Control*. One aim of the two experiments was to compare the effects of testing and restudy practice in authentic test preparation contexts where a certain level of knowledge transfer is required. Given that previous studies were mainly conducted in laboratory settings or focused on the recall of knowledge in classroom settings (e.g., Carpenter et al., 2009; Goossens et al., 2014; Rawson & Dunlosky, 2011; Roediger & Butler, 2011), these experiments can add more evidence on the transfer effects of the two practices on knowledge understanding and application in real assessment contexts.

Besides, participants of the two experiments were recruited from the same cohort in a teacher training college in China, and they were randomly assigned to one of the two experiments following a 6:4 ratio (due to the limited capacity of experimental rooms). The two experiments were identical, except Experiment 1 was conducted two days before the final course test and Experiment 2 was undertaken one week after the test. Thus, students in Experiment 1 had the pressure to prepare for the final course test, while students in Experiment 2 did not have the pressure. By comparing the results across Experiments 1 and

2, it is possible to determine whether the high-stakes context affects the degree of testing effect and restudy effect.

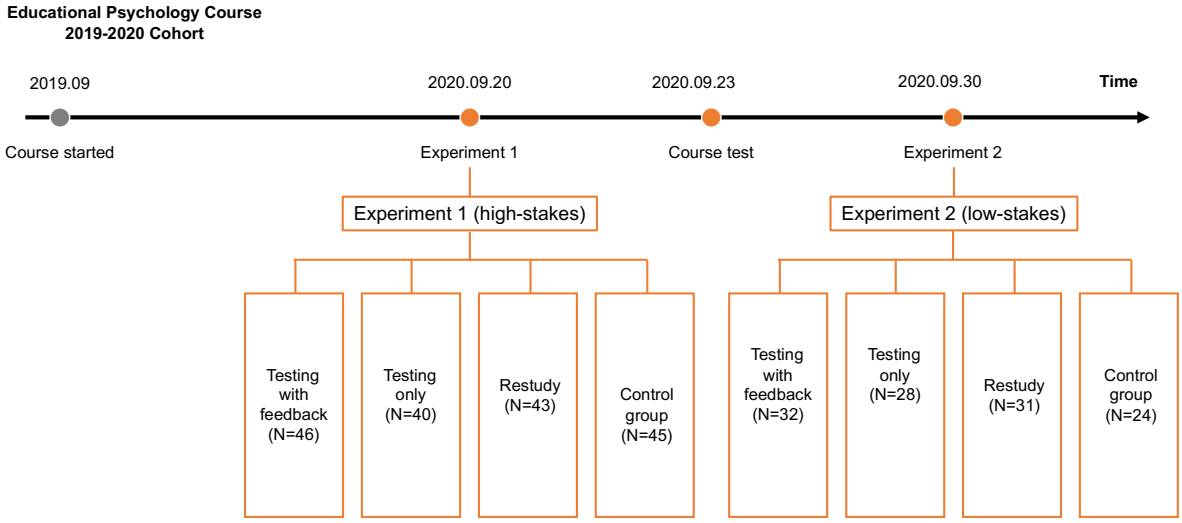


Figure 4 The design of Experiment 1 and Experiment 2.

Experiment 3 and Experiment 4 were quasi-experiments, as students were allocated to experimental and control conditions on a non-random class basis. Students of the two quasi-experiments were from the same cohort, which consisted of eight classes. Three classes participated in Experiment 3, three classes participated in Experiment 4, and two remaining classes served as the business-as-usual control condition (see Figure 5). Although these classes were not equivalent, they were comparable, because students were from the same cohort of students were randomly assigned to classes when they were initially enrolled in the college. Besides, all included classes in the research were given the educational psychology course by the same instructor, thus eliminating the effects of extraneous teacher-related factors. The pretest-posttest design was adopted in the two studies to further eliminate the variance in the test preparation effect caused by students' prior attainment.

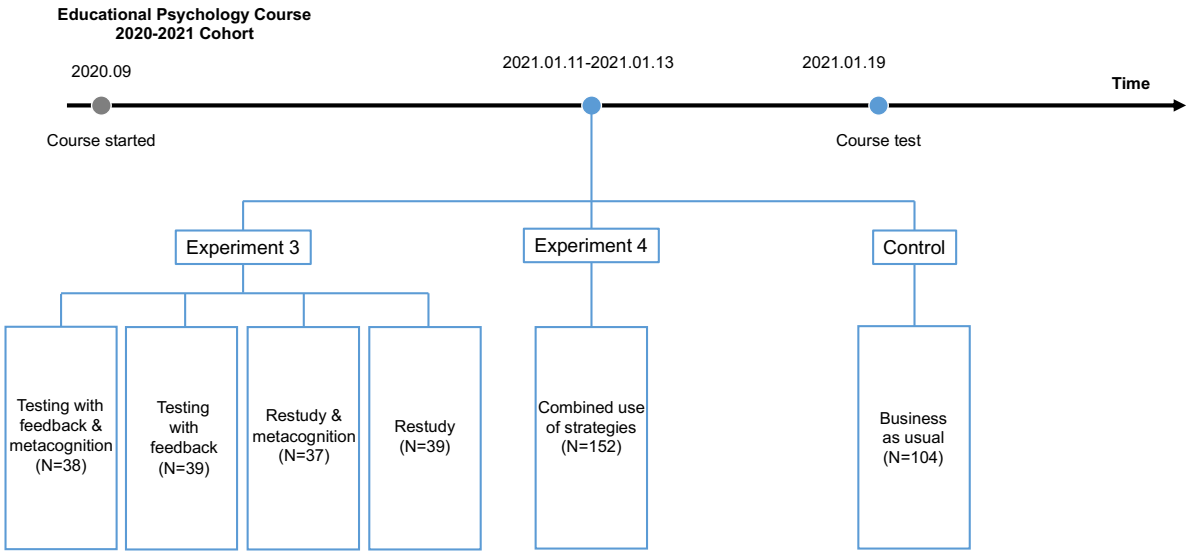


Figure 5 The design of Experiment 3 and Experiment 4.

Experiment 3 partially replicated the design of Experiments 1 and 2 to confirm their findings with a different student cohort. Another purpose of Experiment 3 was to investigate whether students can maximise the testing effect or restudying effect through monitoring and controlling of their test preparation process. Previous studies suggested that metacognitive monitoring and control play an important role in successful learning (Paris & Winograd, 1990; Nelson & Narens, 1994). Accurate monitoring of acquired knowledge can lead to better control of the learning process and thus can lead to better performance (Kostons, van Gog & Paas, 2012; Miller & Geraci, 2014). Ideally, a student who recognises knowledge gaps or difficulties in understanding will invest more time and efforts and implement appropriate learning strategies in order to understand and retain the topic. Whereas students who are confident of their knowledge possession will not engage in further learning, which may result in low performance if the student does not actually master the knowledge. Händel et al.'s (2020) study suggested that providing students with metacognitive training and monitoring and control tasks significantly added to the testing effect. Therefore, based on findings of previous literature, a new variable, metacognition, was added in Experiment 3,

resulting in a two (strategy use for test preparation: testing vs. restudy) by two (metacognition: yes vs. no) factorial design. Note that to simplify the aliases of experimental groups, 'metacognition' was used in the group aliases in this thesis to indicate the two groups receiving a brief training on metacognition and doing metacognitive monitoring and control tasks during the intervention. This design allowed the comparison between the effects of testing and restudy, and an examination of whether students' metacognitive monitoring and control of their usage of testing and restudy could improve the effectiveness of the preparation. Possible interactions, between metacognitive tasks and test preparation strategies, were examined, to provide more precise inferences regarding the conditions under which test preparation effect could be boosted.

As the first three experiments manipulated test preparation conditions, individual participants were required to use one specific strategy to prepare for the test during the intervention. However, in real test preparation scenarios, students may employ both strategies and use them for different purposes. Some previous studies suggested that students may use restudying earlier in their learning process and testing later (Janes et al., 2018; Kornell & Soon, 2009). And other studies found that students used testing as a way to diagnose their learning, so they tended to choose testing practice more often when they predicted that the item would be easy (Toppino et al., 2018). To better understand students' use of the two strategies, Experiment 4 sought to clarify the circumstances under which a specific strategy might be preferred during students' self-study for test preparation. Three main factors that may affect strategy use were examined, including perceived item difficulty, task nature, and self-monitoring of learning. Following an explanatory-sequential approach, this study also collected students' perceptions on the effectiveness of the two strategies. These qualitative data were used for the subsequent interpretation and clarification of the quantitative data on students' strategy choices. In addition, students' performance on pre-test, post-test, and the

final course test that was held one week after the experiment were also analysed to examine the effect of students' combined use of the two test preparation strategies. The effect was also compared with effects of using single strategies in Experiment 3, to draw conclusions on whether combined use of strategies leads to greater gains.

3.2 Overall methodology of the four experimental studies

3.2.1 Research site

The four (quasi-) experiments were conducted in a teacher training college in China. This college was chosen because it was the only institution that allowed me to conduct on-site data collection during the COVID-19. Three other institutions were also contacted but they were closed to the public due to the local pandemic restrictions. This teacher training college is located in the southwest of China and provides vocational upper secondary education. Students enter in the school around the age of 15 after graduating from lower secondary schools. This college provides several five-year long programmes that prepare students for careers as subject teachers in preschools and primary schools, offering a wide range of courses to develop students' subject knowledge as well as pedagogical and other professional skills.

It is worth noting that students in college were not typical high school students. In China, about 50.8% of lower secondary graduates enter high schools, while about 23.5 % of them enter vocational schools (MoE, 2021). Specialised teacher training schools make up a relatively small proportion of all secondary vocational schools. Taking the province in which the research was conducted as an example, only four out of a total of 257 secondary vocational schools are teacher training schools. However, because of the relatively high employment rate of graduates from teacher training schools and the social recognition of the teaching profession, the programmes provided by the college are usually competitive and

students get admitted with relatively higher Zhongkao (i.e., the high school entrance exam) scores, which can be approximately equivalent to the admission scores of regular high schools. Besides, due to the Free Teacher Education (FTE) policy, about 30%-40% of students enrolled in this school receive a waiver of tuition fees, based on their outstanding Zhongkao performance (usually equivalent to admission scores to prominent high schools). According to the policy, FTE pre-service teachers are required to pass the National Teacher Certification Examination (NTCE) and work in rural schools. They will be given priority when they apply for jobs.

To start a career as a teacher, students usually take the NTCE from the beginning of their fourth year, and about half of them repeat the examination multiple times before they get the certification. The NTCE consists of two parts: paper and pencil tests, and an interview. Educational psychology is a compulsory sub-test in the examination and a compulsory course for all third-year students in the college. Due to the importance of this subject, this college invited me to design intervention programmes to prepare students for the final course tests of the educational psychology courses. Therefore, participants of the four (quasi-) experimental studies had some prior knowledge of cognitive and metacognitive learning strategies (e.g., retrieval practice, monitoring and control, etc.), as effects of these strategies had been taught as a part of the course syllabus.

The educational psychology course lasts for one-year, and students are required to pass two course tests in order to get the course credits, one test in the middle of the academic year, usually in January, and the other one at the end of the academic year, usually in June. However, due to the outbreak of the COVID-19, students enrolled in the course in 2019-2020 academic year took the class online from February 2020, and the course test was postponed to mid-September 2020, at which time the school was re-open, and the new academic term was about to start.

3.2.2 Participants

Participants of the four (quasi-) experiments were third-year students from the teacher training college. The first two experiments invited the cohort enrolled in the educational psychology course in 2019-2020 academic year, while the last two experiments invited the cohort enrolled in the course in 2020-2021 academic year. The course was taught by the same instructor for the two years. Participants took part in the study in exchange for course credits. Table 1 presents the descriptive information of number, age and gender of participants in each experiment.

Table 1 Information of participants in four experiments.

Experiment number	**Number of participants**		**Age**	**Gender**
	Overall	**Number in each group**	**Mean (SD)**	**(percentage of female)**
Experiment 1	172	Testing&Feedback = 46 Testing only = 40 Restudy = 43 Control = 45	18.3 (.85)	68%
Experiment 2	115	Testing&Feedback = 32 Testing only = 28 Restudy = 31 Control = 24	18.3 (.83)	70%
Experiment 3	257	Testing & feedback = 39 Testing & feedback plus metacognition = 38 Restudy = 39 Restudy plus metacognition = 37 Control = 104	17.34 (1.21)	72%
Experiment 4	256	Experimental group = 152	17.51 (1.14)	65%

	Control = 104

Note: The control group specified in Experiment 3 and Experiment 4 were the same.

3.2.3 Design

All four (quasi-) experiments employed the pretest-posttest control and experimental group design. Experiments 1 and 2 were true experiments; the former was conducted two days before the high-stakes course test while the latter was conducted one week after the course test. Participants from the same cohort were randomly assigned to the two experiments, making it possible to compare the effect of test preparation between the high-stakes context (Experiment 1) and the low-stakes context (Experiment 2).

Additionally, in each experiment, participants were randomly assigned to one of the four groups: *Testing with feedback*, *Testing only*, *Restudy*, and *Control*. The difference in participants' performance between these four groups would suggest how different test preparation strategies had varying effects on test results. In Experiment 1, participants' performance on the final course test was also analysed to assess the effect of test preparation on the long-term retention, whereas Experiment 2 did not include any delayed measurement of participants' performance.

Experiment 3 and Experiment 4 used the non-equivalent group pretest-posttest design. It was not possible to employ a true experimental design in the two studies for practical considerations. At the time of these two experiments, the school was unable to provide enough empty classrooms for the experiments. Thus, the experiments were carried out in the educational psychology classroom. Since each class took the course at different times, the two experiments were spread over three days. Students from eight classes were invited to participate in the two experiments. Three classes participated in Experiment 3. As shown in Table 2, Experiment 3 used a two (test preparation strategies) by two (metacognitive monitoring and control tasks) factorial design, resulting in four experimental groups: *Testing*

with feedback, *Testing with feedback plus metacognition*, *Restudy*, and *Restudy plus metacognition*. Participants from each class were randomly assigned to one of the four experimental groups. In the two metacognition groups, except using the assigned strategy to review educational psychology topics, participants also completed a series of monitoring and control tasks, including rating the difficulty of the topic, rating their confidence on knowledge mastery of the topic, and making decisions on repeated review the current topic or proceed to the next topic. Participants' ratings and study decisions were recorded for the analysis.

Table 2 Design of Experiment 3.

		Metacognitive training and monitoring and control tasks	
		Provided	**Not provided**
Test preparation strategies	**Testing with feedback**	Testing with feedback plus metacognition	Testing with feedback
	Restudy	Restudy plus metacognition	Restudy

Participants from other three classes took part in Experiment 4, where they were allowed to blend the use of testing and restudy to learn educational psychology topics. During the intervention, participants in Experiment 4 chose either testing or restudy to review each educational psychology topic. Besides, they were also required to complete the monitoring and control tasks as what participants in metacognition groups in Experiment 3 did. Two

remaining classes served as the control condition for Experiments 3 and 4, where they watched a video on teacher professional development instead of doing test preparation practices. In the two experiments, participants' performance on the final course test was also analysed to assess the effect of interventions on the long-term retention.

3.2.4 Materials

The materials used in the experiments mainly consist of intervention materials and test materials. All materials used in the four experiments were from authentic curriculum material and past test papers. Materials were provided in Chinese. All materials used in Experiments 1 and 2 were the same, and materials used in Experiments 3 and 4 were the same. By the time Experiments 1 and 2 were conducted, students had already completed the entire course, while those in the Experiments 3 and 4 had only completed half of the course. Therefore, materials used in the first two experiments covered a wider range of educational psychology topics than those used in the latter two experiments. The specific use of materials in the four experiments is detailed below.

Intervention Materials

The one-year educational psychology course covered eight chapters related to educational psychology theories in two academic terms, with four chapters taught in each term. As Experiment 1 and Experiment 2 were conducted at the end of the entire course, the interventions in both experiments focused upon topics from six chapters in the curriculum, which had been highlighted in instruction and emphasised by the final course test. In total, 24 topics were chosen from the six chapters of the curriculum: 'Psychological development and education of primary school students', 'Learning and learning theories', 'Learning strategies', 'Knowledge and skills', 'Morality', and 'Teachers' Professional Development'.

In contrast, interventions in Experiment 3 and Experiment 4 covered a relatively smaller range of topics, as students just completed half of the curriculum. The interventions focused upon 16 topics chosen from four chapters ‘Psychological development and education of primary school students’, ‘Learning and learning theories’, ‘Learning strategies’, and ‘Knowledge and skills’.

For each topic, one statement and one testing question were created by summarising the key information about the topic based on the textbook and teachers teaching slides. The researcher and the course teacher discussed each statement and each testing question to make sure they were stated properly and accurately. Examples of testing questions and statements are presented in Table 3. Participants could access the intervention materials through the PsyToolkit platform (Stoet, G. 2010, 2017). Each statement or test item was presented on one separate page, and the presentation order of concepts was consistent.

Additionally, participants in the two metacognition groups in Experiment 3 and all participants in Experiment 4 received 20 minutes of instruction on the important role that metacognition plays in students’ learning. Students watched a video and practised sample monitoring and control tasks. During the intervention, these students did a series of metacognitive monitoring and control tasks, including rating the difficulty of the topic, rating confidence before and after learning each topic, and making decisions on repeated studies. To be specific, students were asked to rate the difficulty of each topic on a 10-point Likert-type scale ranging from 1 (very easy) to 10 (very difficult). Before and after reviewing the topic (no matter what strategy they used), students were required to rate their confidence in the knowledge mastery of the topic from 1 (very unsure I have mastered the topic) to 10 (very sure I have mastered the concept). After each time students reviewed the topic and made the ratings, a control task was followed asking students to decide whether they want to learn the next topic or revisit the current topic.

Table 3 Examples of practice questions and statements.

Concepts	Testing groups	Restudy group
Zone of proximal development	**Question**: Please explain the concept of ZPD. **Feedback**: ZPD is the difference between what a child can do alone and what child can do with the help of a teacher, includes a child's emerging skills.	ZPD: the difference between what a child can do alone and what child can do with the help of a teacher, includes a child's emerging skills.
Piaget's cognitive development theory	**Question:** A child named Ming can deduce the relationship between A, B and C based on the principle that A is bigger than B, B is bigger than C, then A is bigger than C. For example, 'Zhang is taller than Li, Li is taller than Wang, so Zhang is the highest.' According to Piaget's cognitive development theory, Ming is at _____ stage. A. Sensorimotor stage B. Preoperational stage C. Concrete operational stage D. Formal operational stage **Feedback**: The correct answer is D. In the formal operational stage (age 12 and up), children tend to reason in a more abstract, systematic, and reflective	According to Piaget's cognitive development theory, in the formal operational stage (age 12 and up), children tend to reason in a more abstract, systematic, and reflective way. They are more likely to use logic to reason out the possible consequences of each action before carrying it out. Ming can deduce the relationship between A, B and C based on the principle that A is bigger than B, B is bigger than C, then A is bigger than C. For example, 'Xiao Zhang is taller than Xiao Li, Xiao Li is taller than Xiao Wang, so Xiao Zhang is the highest.' According to Piaget's cognitive development theory, Ming is at Formal operational stage.

	way. They are more likely to use logic to reason out the possible consequences of each action before carrying it out.	
Learning strategies	**Question:** What is the difference between cognitive strategies and metacognitive strategies? Please give a brief description with examples. **Feedback**: Cognitive strategies are specific methods/techniques of information processing, which help effectively solve problems or complete tasks. Cognitive strategies include repetition, organizing and summarizing strategies. For example, copying, taking notes, making outlines, making associations, etc. Metacognitive strategy is the cognitive strategy students used on their cognitive process, including the monitor and control strategy of their own cognitive process, which helps students effectively arrange and regulate the learning process. Metacognitive strategies include planning strategies, monitoring strategies and adjustment strategies. For example, setting learning goals, paying attention, taking self-tests, etc.	Cognitive strategies are specific methods/techniques of information processing, which help effectively solve problems or complete tasks. Cognitive strategies include repetition, organizing and summarizing strategies. For example, copying, taking notes, making outlines, making associations, etc. Metacognitive strategy is the cognitive strategy students used on their cognitive process, including the monitor and control strategy of their own cognitive process, which helps students effectively arrange and regulate the learning process. Metacognitive strategies include planning strategies, monitoring strategies and adjustment strategies. For example, setting learning goals, paying attention, taking self-tests, etc.

Test materials

Participants in all four experiments had to complete a pre-test and a post-test. In each experiment, two parallel test papers were used for the pre-test and post-test, both of which were designed in a similar structure to the course test. Experiments 1 and 2 used the same set of the two test papers—*Test paper A* and *Test paper B*, which difficulty levels were likely to be equivalent based on students' performance in the pre-test, t (285) = -0.75, p = .455. Each test paper contained 28 items in total, 20 of which were multiple-choice questions (2 points for each), 4 of which were definition recall questions (6 points for each) and 4 of which were case analysis questions (9 points for each). The test structure of test paper A and B is illustrated in Table 4. Experiments 3 and 4 used another set of two parallel papers—*Test paper C* and *Test paper D*, which were shorter than tests used in prior two experiments due to the constraints of course time. The difficulty levels of the two papers were similar based on students' performance in the pre-test, t (407) = 0.44, p = .651. As shown in Table 4, each test paper contained 10 multiple-choice questions, 2 definition recall questions, and 4 critical analysis questions. Students in Experiments 3 and 4 were also required to rate their confidence on a 10-point scale in each test item during the pre- and post-tests.

All test items were developed by the course teacher based on previous test papers and teaching slides. Knowledge assessed in these tests were covered in the intervention sessions, but they were presented in different ways. For example, during the intervention, participants in experimental groups learned the theory of *Discovery Learning* proposed by *Bruner* by either answering multiple-choice questions (testing groups) or by restudying the statement of key points (restudy groups), while this concept was tested in one of the case analysis questions by asking students to evaluate this theory critically and talk about what they have learned from the theory as the pre-service teacher. In this way, the design simulated educational assessment practices in the real world, where students could study the content

that will be tested, but in real tests, the content is usually presented in a way that cannot be exactly predicted and will not be always identical to what they have learned.

In addition, students in Experiment 1, Experiment 3 and Experiment 4 took the final course test several days after the experiments. The course tests were developed by two teachers in the school and lasted for 90 minutes. The course test taken after Experiment 1 assessed a wider range of knowledge than the coverage of intervention. To ensure that the analysis of final course test performance can indicate the long-term effect of intervention, only those test questions that assessed topics covered by the intervention were selected and scored for the analysis, resulting in a total of 83 points. Since the pre- and post-tests were based on a total of 100 points, raw scores on the course test were converted to a score out of 100 to allow for the calculation of score gains between the pre-test and the course test. The score conversion was done by calculating the percentage of each participant's raw score out of the total score and converting it to a score out of 100.

Regarding the course test after Experiments 3 and 4, it was based on the same four chapters that were the focus of the intervention in the two experiments, and topics assessed in this course test were covered by the intervention. Thus, students' raw scores were used in the analysis directly to indicate the long-term retention of test preparation effects derived from interventions.

Table 4 The structure of the pre/post-tests used in four (quasi-) experiments.

Components	**Quantity of items and points**	
	Test paper A & B	**Test paper C & D**
Multiple-Choice Questions	20 items, 2 points for each	10 items, 3 points for each
Short-Response Questions (Definition recall questions)	4 items, 6 points for each	2 items, 12 points for each
Short-Response Questions	4 items, 9 points for each	4 items, 11 or 12 points for each

(Case analysis questions)		
Overall	28 items, 100 points in total	16 items, 100 points in total

3.2.5 Procedure

The Pilot

Three weeks prior to carrying out the studies, the researcher invited ten students who enrolled in the educational psychology course in previous years to attend a pilot study. The purposes of the pilot were to check if there were mistakes and unclear instructions in the materials, and to make sure the online PsyToolkit platform (Stoet, G. 2010, 2017) functioned well. Pilot participants reported that the platform was simple to use and that the instructions and tasks presented on it were clear. However, some participants pointed out that they failed to load pictures inserted in the online platform because of poor network connection. Besides, one participant suggested to set a minimum time for reading each statement so that participants did not rush through the pages and skim over topics. Based on participants' feedback, the researcher removed all pictures from the materials and kept the texts only to make sure that materials could load quickly. In addition, a minimum time block of 5 seconds was set for each statement for the restudy group. Following the pilot, the instructions for the use of the platform were updated, which were delivered to all participants in the experiments.

Experiment 1 and Experiment 2

Information sheets and consent forms were distributed to all participants one day before conducting the experiments. On the day of the experiments, students who participated in the studies had the chance to sign and submit the consent form. Later, a 10-minute training session on how to use the PsyToolkit platform was provided. To familiarise participants with the platform, sample materials which were developed based on the course content were provided to participants to practice but were not included in the experiment. After the

training, participants were informed of which group they were allocated to, as well as the procedures of the experiment and a question-and-answer session followed to ensure participants' queries were addressed.

Experiment 1 was conducted two days prior to the final course test. Participants went to computer rooms to take the experiment. Each computer room was used for a certain group. Two equivalently difficult test papers (*Test paper A* and *Test paper B*) were used (see section 3.2.4) for the pre-test and post-test and were administered in a paper-and-pencil format. Half of the students in each experimental condition were randomly assigned to use Test A for the pre-test and Test B for the post-test, while the other half was randomly selected to use Test B for the pre-test and Test A for the post-test. The pre-test lasted for 40 minutes and students had a 10-minute break between the pre-test and the intervention. The intervention was self-paced, and the maximum time for it was 60 minutes. Participants in the experimental groups were allowed to take notes using a pencil and paper during the intervention. Meanwhile, the control group solved sudoku puzzles, and they were allowed to discuss and share experiences with each other. A post-test was provided after all participants had finished the intervention.

After the experiment, all testing and restudying materials were distributed to participants including those in the control group. The course teacher reiterated that participants could use them as additional materials to prepare for the final course test. The final test was held two days after Experiment 1.

Experiment 2 was conducted one week after the course test. The procedures of this experiment were identical to the ones adopted in the first experiment.

Experiment 3 and Experiment 4

The two experiments were conducted concurrently in classrooms during course time. Before the experiments, information sheets and consent forms were distributed to every student. The researcher requested that test materials and training materials should not be

disclosed to others. After receiving the signed consent forms, the researcher assigned participants to different groups, and informed participants of which group they were placed in two days before the experiment. A day before the experiment, participants assigned to the same group attended a training session together. The course teacher and the researcher provided instructions on the use of the PsyToolkit platform to all participants in experimental groups. Participants in the metacognition groups in Experiment 3 and all participants in Experiment 4 also received instructions on the importance of metacognition, as well as how to do metacognitive monitoring and control tasks. Participants practiced with several example questions to familiarise themselves with the tasks. A question-and-answer session was then provided to each group after the training to help participants further understand the experiments' procedures.

After the training session, the pre-test was administered in a paper-and-pencil format, using two equivalently difficult test papers (*Test paper C* and *Test paper D*). Half of the participants were randomly assigned to use Test paper C for the pre-test, and the other random half participants used Test D for the pre-test. The pre-test lasted for 25 minutes, and the intervention started immediately after the pre-test. The intervention was presented on the PsyToolkit platform and participants could access it via either laptops, pads or mobile phones. Participants were encouraged to use whatever they were comfortable with, and they were also given the option to use paper and pencil to facilitate their learning (for example, taking notes). Before each experiment, the researcher mentioned to the participants that they were expected to complete the study of all topics within 60 minutes. The practice was self-paced, and the research reminded participants of the remaining time every ten minutes.

During the intervention in Experiment 3, participants in *Testing with feedback* and *Restudy* groups reviewed educational psychology topics by either testing or restudy. Participants in the two metacognition groups—*Testing with feedback plus metacognition* and

Restudy plus metacognition—were asked to rate the difficulty of each topic prior to reviewing it, and also rated their knowledge mastery before and after reviewing each topic. After reviewing each topic, participants were given the option to proceed to the next topic or review the current topic once again. The intervention process of Experiment 3 is illustrated in Figure 6.

During the intervention in Experiment 4, each topic of the to-be-learned concept was presented on the screen first, after which participants rated the difficulty of the topic and their confidence of knowledge mastery. Then they decided which strategy to use to learn this topic, choosing from restudying the statement or answering a practice question. The type of the practice question (multiple-choice or short response) was provided on screen. After reviewing each topic, participants were required to rate their confidence of mastery of this topic again. Then, they decided whether to review the same topic again or move to the next topic. Students made as many attempts as they wanted on each topic, choosing from restudying or testing for each attempt. The intervention process of Experiment 4 is illustrated in Figure 7.

After the interventions in Experiments 3 and 4, students took a 10-minute break before taking the post-test, which was conducted in the same manner as the pre-test. Students who used Test C in the pre-test were allocated Test D for the post-test and vice versa.

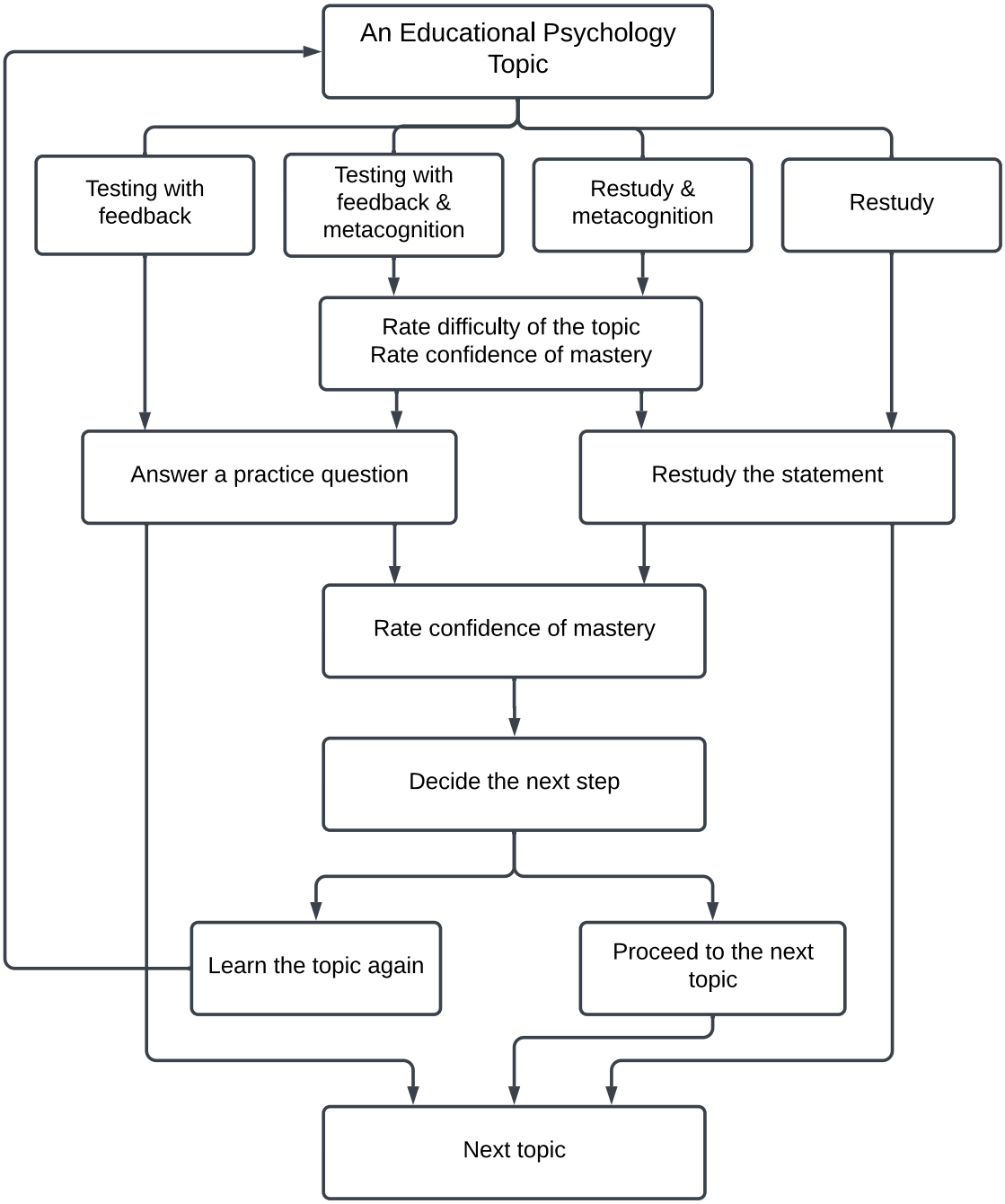


Figure 6 The Intervention design of Experiment 3.

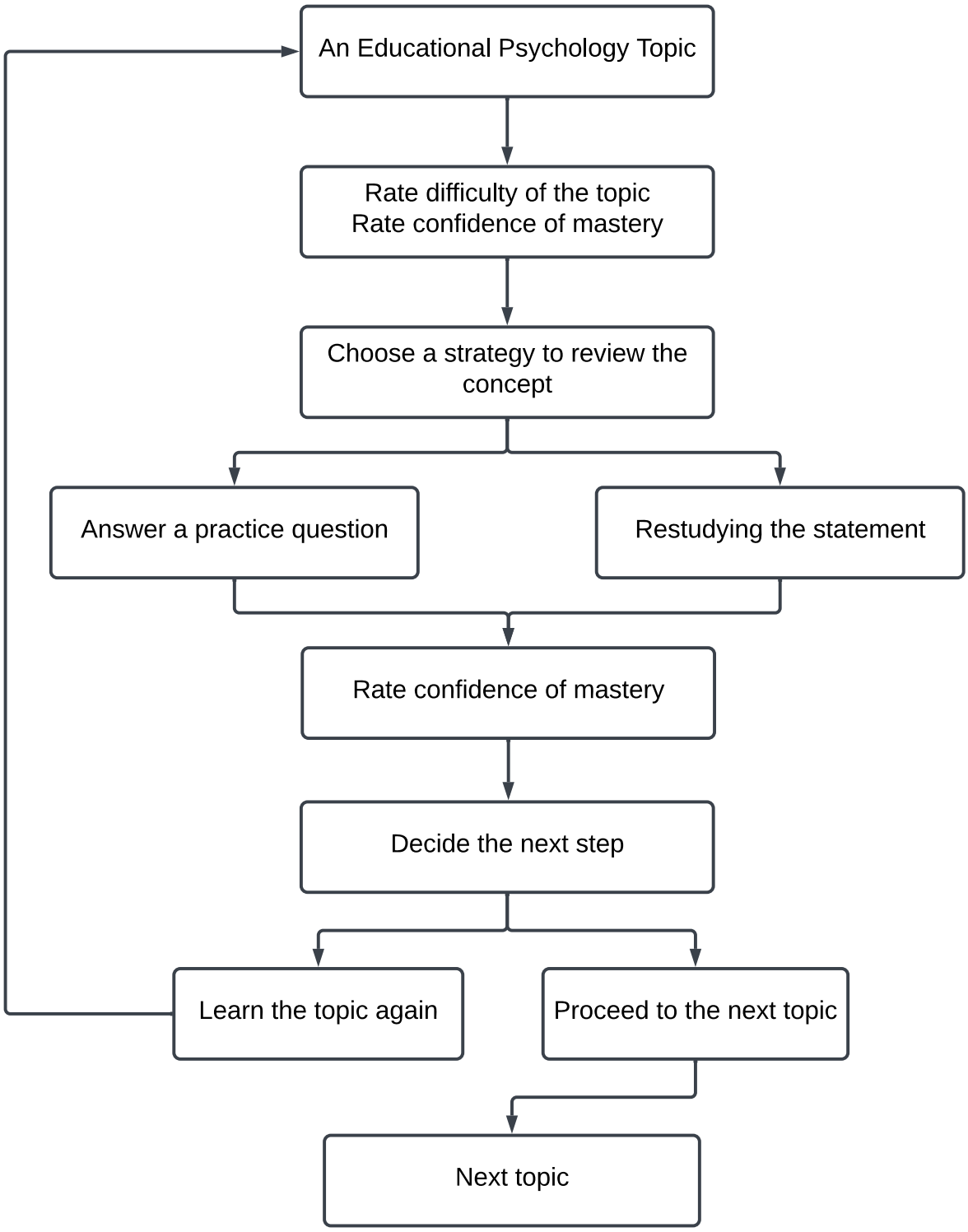


Figure 7 The intervention design of Experiment 4.

3.2.6 Analytical strategy for the experiments

The analytical strategies for each experiment and for the cross-experiment investigation are summarised in this subsection. The main analyses conducted in Paper 2 and Paper 3 are presented first, followed by a snapshot of four sets of additional analyses in Chapter 7.

Analyses in research papers

Effects of test preparation strategies

In each of the first three experiments, a one-way analysis of covariance (ANCOVA) was adopted to compare the difference in post-test scores between test preparation groups when students' pre-test scores were controlled as the covariate. Post hoc Bonferroni tests were used to indicate the pairwise comparisons between groups. Results of these ANCOVA models shed light on the effect of different test preparation strategies on test performance. In Experiment 1 and Experiment 3, one-way ANOVA models were also used to compare the group difference in course test performance, using pre-test performance as the covariate. Results of these analyses can suggest whether the effect of test preparation practices retained for a long-term.

As the test preparation interventions in the three experiments were self-paced, differences in the amount of time students in different groups spent during the intervention are likely to influence the effect. To control this possible confounding variable, in each of the first three experiments, another ANCOVA model was employed using experimental conditions as the independent variable, pre-test performance and time use as the two covariates. The ANCOVA models compared the group difference in post-test scores (for Experiment 1, Experiment 2, and Experiment 3) and in delayed course test scores (for Experiment 1 and Experiment 3) across experimental conditions in every experiment.

Impact of the level of stakes of the setting

To understand whether the test preparation effect varies under high-stakes and low-stakes test conditions, a cross-experimental comparison was further conducted between Experiment 1 and Experiment 2. A two (experiment sequence: Experiment 1 *vs*. Experiment 2) by four (test preparation condition: Testing with feedback *vs*. Testing only *vs.* Restudy *vs.* Control) mixed factorial ANCOVA on post-test performance was conducted, using pre-test performance as the covariate, to compare the test preparation effects across the two experiments (different stakes of the settings).

Study behaviours

Students' study behaviours, represented as time use and study attempts, were also analysed based on data from Experiments 3 and 4. First, descriptive statistics for students' study behaviours were generated. Then, generalised linear mixed-effects models using the glmer function in the 'lmer4' package (Version 1.1–21; Bates, Maechler, Bolker, & Walker, 2015) in the R Studio (R Core Team, 2018) were conducted to investigate whether and how students' perceived difficulty of topics and their level of confidence predicted their decisions on whether to repeat their study attempts in the two experiments respectively. The analytical models in Experiment 4 also investigated whether and how students' strategy choices among testing and restudy were affected by their judgements and topic-specific characteristics (i.e., whether the test item was a multiple-choice question or a short-response question if students chose to test themselves on the topic). The formulation of every generalised linear mixed-effects model can be found in Paper 2 and Paper 3.

Students' perceptions of strategies

As introduced in earlier sections in this chapter, some qualitative data were collected after the intervention in Experiment 4 to understand students' perceptions of testing and restudy strategies, as well as to help interpret the data on their actual behaviours during the

test preparation session. The qualitative data were analysed using thematic analysis, following procedures suggested by Braun and Clarke (2006). The course teacher was invited to assist in coding data. The original transcripts of students' responses (in Chinese) were coded, as both the course teacher and the researcher were native Chinese speakers. In the beginning of the coding, the researcher and the course teacher coded the responses of 30 students (20%) separately, and they agreed with each other for over 85% of the shared coding data. Disagreements were reconciled through discussion. Then, the researcher read all remaining students' responses, applying the codes to label data extracts. Recurrent codes and possible themes were identified by the researcher during coding. All defined themes were presented to the course teacher and were discussed to reach agreement between the two coders. The finalised themes and codes were translated into English by the researcher to be presented and discussed in this thesis.

Additional analyses

The effect of combined use of strategies

As the effect of the combined use of strategies (Experiment 4) was not examined in Paper 3, the analysis was presented in Chapter 7. The main purpose of this analysis was to investigate whether the effect of using testing and restudy in combination exceeded the effects of using single strategies. Therefore, a series of one-way ANCOVA models were conducted to compare test results in Experiment 4 and Experiment 3. The dependent variables were the post-test and the course test performance, and the independent variable was the test preparation conditions across Experiments 3 and 4 (*Testing with feedback*, *Testing with feedback plus metacognition*, *Restudy*, *Restudy plus metacognition*, *Combined use of strategies*, and *Control).* Students' performance on the pre-test was used as a covariate.

Test preparation effects across four (quasi-) experiments

A continuously cumulating meta-analysis of test preparation effect among the four (quasi-) experiments in this research was performed, so as to provide an estimate of the overall test preparation effect on the test performance in the educational psychology course. As test preparation conditions were nested within experiments, a correlated and hierarchical random effects model was used. Four variables - *test preparation strategy*, *inclusion of metacognitive training and tasks*, *stakes of tests*, and *time interval between test preparation and the test* – were coded based on the design of each experiment. The four variables were further analysed by subgroup analysis and meta-regression to explore if they were significant moderators for the test preparation effect.

Relationships between prior attainment and score gains

The relationship between students' prior attainment level and their score gains after the experimental intervention was analysed to extend the understanding of whether and how individual differences affect the effect of test preparation. In order to compare the benefits of intervention for students with different prior attainment levels, score gains were calculated and used as the dependent variables. As the pre- and post-tests and the course tests had the same overall score, score gains in the post-tests and final course tests were calculated by deducting the pre-test scores from them. The prior attainment level was the independent variable, and it was classified in two approaches.

Approach 1: The absolute approach.

Experiments 1 and 2 recruited participants from the same cohort of students and used a same set of materials, and Experiments 3 and 4 recruited participants from another cohort and used another set of materials. Therefore, students in Experiments 1 and 2 were grouped together and partitioned into three attainment levels based on whether their pre-test scores fell within the upper, middle, or lower one third. Similarly, students in Experiments 3 and 4 were

grouped together and classified into three levels by ranking their pre-test scores. Note that there were few low prior attainment performers in Experiment 2 due to the absolute way in which their performances were classified (across experiments). A series of ANOVA models were performed to compare the group difference in score gains across different prior attainment levels and different test preparation conditions.

Approach 2: The relative approach.

The relative approach was experiment-specific, that students in each experiment were ranked based on their pre-test performance and divided into three levels. Note that there were more equal numbers in the low, middle and high groups *within* each experiment, due to the relative way in which the classifications were conducted. A series of ANOVA models were used within each experiment to compare different score gains between students with different prior attainment levels by different test preparation groups.

Relationships between prior attainment levels, test preparation behaviours and monitoring accuracy

Students' study behaviours during the online test preparation sessions were analysed in relation to their prior attainment levels. Study behaviours were reflected in four variables: time uses, the number of overall study attempts, and number of testing and restudy were selected respectively. A series of ANOVA models were used to compare group differences on these variables across absolute prior attainment level and test preparation conditions. The prior attainment level in the absolute sense was used because it enabled meaningful comparisons across experiments.

Besides, in Experiments 3 and 4, students rated their confidence in each test item during the pre- and post-tests. Therefore, their absolute accuracy of their monitoring was calculated, using the formula shown below (Schraw, 2009):

$$\text{Absolute Accuracy Index} = \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i)^2$$

where c_i corresponds to a confidence judgement, p_i corresponds to a performance score, and *N* is the number of rated items. To note, performance scores for multiple choice items were converted to either 1 (incorrect) or 10 (correct), while scores for short-response items were converted to a 10-point scale based on the percentage of scores earned out of the total score for each item. A two-way ANOVA model was used to compare group differences on absolute monitoring accuracy across absolute prior attainment levels and test preparation conditions.

3.3 Ethical considerations

The (quasi-) experimental studies in this research received ethical approval through the Central University Research Ethics Committee (CUREC) at the University of Oxford (Appendix A and Appendix B).

The research complied with the ethical guidelines of BERA (2011) which outlines five primary principles of ethical research: minimizing harm; respecting autonomy; protecting privacy; offering reciprocity; and treating people equitably. In this research, all studies aimed to limit any inconvenience to the participants and the college. The implementation of each study was discussed thoroughly with the course teacher, class headteachers and the school headmaster to make sure possible inconvenience would be minimised. The first two experiments were conducted in common computer rooms in a separate building to make sure that students who did not participate in the study would not be bothered. When the Experiments 3 and 4 were conducted, the common computer rooms were not available, so the two experiments were conducted during the time of the course in classrooms. Course teachers in the college usually provide their classes with a test preparation session before the final test, and the teacher who assisted this study was willing to take the experimental interventions as

the replacement of these sessions. The teacher thought the experiment was not only beneficial for students to improve test performance, but also for him to understand the effectiveness of different test preparation strategies and improve his teaching. Thus, the study did not disrupt teaching.

Participants' genuine willingness to participate in the research was considered when conducting the study. Therefore, an opt-in strategy was adopted. Students were informed that their participation was voluntary and that they could quit anytime. Besides, the anonymity of participants was always preserved in the study. Students used nicknames and student numbers to participate in the experiment. Their student numbers were used to match their performance in the experiment and the final course test. All the data were anonymised and extra attention was paid to keep their data confidential, including storing data files on a password-secured laptop with the online folder which stored the data being only accessible to the researcher and her supervisors.

All participants, whether in experimental groups or control groups, received course credits for their participation. More importantly, mindful of the possibility that students in control groups may underperform in the final course test because they did not receive the test preparation intervention, all materials used during the intervention were provided to them after the experiment for their self-study.

Part II
Research Papers

Chapter 4. Paper 1 81

Chapter 5. Paper 2 139

Chapter 6. Paper 3 190

Chapter 4. Paper 1

This chapter presents the first research paper (Paper 1), which is a meta-analysis of test preparation effects on educational tests. This meta-analysis mainly concerned the effect on large-scale educational tests, as the main test preparation literature is developed based on discussions around the large-scale tests, most notably college admission tests and English tests. Thus, relevant data are more substantial and accessible. Possible moderators of test preparation effect were discussed based on the sub-group analysis and meta-regression results. The results demonstrated that the breadth of test preparation intervention, the use of workbook, instructions on test-taking skills and test management skills were significant moderators of the effect. Besides, interventions on English as a Foreign Language (EFL) had a significantly greater average effect size on test performance than the effect that interventions on verbal tests had. However, the use of practice tests and sample items did not moderate the effects of test preparation interventions, which finding appears to be inconsistent with findings reported in previous literature.

The Impact of Test Preparation on Performance of Large-scale Educational Tests: A Meta-analysis of (Quasi) Experimental Studies

Zhanxin Hao[1]; Jo-Anne Baird[1]; Yasmine El Masri[2]; Kit Double[3]

[1] Department of Education, University of Oxford, Oxford, UK

[2] Ofqual, Coventry, UK

[3] Faculty of Science, School of Psychology, The University of Sydney, Sydney, Australia

Corresponding details: zhanxin.hao@education.ox.ac.uk

Funding: This work was supported by Chinese Scholarship Council [grant number: 201808060181].

The Impact of Test Preparation on Performance of Large-scale Educational Tests: A Meta-analysis of (Quasi) Experimental Studies

Abstract

Test preparation has been widely used to optimise test-takers' potential on large-scale educational tests, meanwhile, people questioned its effect and raised concerns that it may narrow students' learning. This meta-analysis provided a statistical synthesis of existing experimental and quasi-experimental studies on effects of test preparation interventions on students' test performance on large-scale educational tests. Results from 30 included studies suggested that students' test performance can be significantly improved by test preparation compared with no preparation ($g = 0.38$, $p < .001$). The effect of test preparation could be attributed to an increase of domain-specific knowledge or test-specific cognitive skills. Using workbooks, developing students' test preparation management skills and teaching test-taking skills were found to be significant moderators of test preparation effect. Besides, we also found that students who had previous experience of taking the test or attending relevant test preparation benefited from test preparation more than those who had not. Little evidence of practice effect (i.e., using sample items and practice tests) on test preparation was found in the current study.

Keywords: test preparation, test preparation effect, test performance, large-scale educational tests, meta-analysis

The Impact of Test Preparation on Performance of Large-scale Educational Tests: A Meta-analysis of (Quasi) Experimental Studies

1. Introduction

Large-scale educational tests are an inherent part of education systems throughout the world. Test scores have been used tremendously by educational authorities internationally and nationally for making decisions related to gatekeeping, accountability, and policymaking. The increasing power of test scores has spawned an ever-growing demand for test preparation inside or outside schools aimed at helping students to prepare for tests (Bray, 2007; Bunchmann et al., 2010, Ross, 2008). Some schools provide students with test preparation courses to achieve greater accountability. Besides, a large number of commercial educational institutions built well-established test preparation training systems which not only provide group courses and private tutorials, but also publish practice books or even develop online training systems. The underlying rationale for these test preparation activities is that test preparation is conducive to maximizing test results.

Although the test preparation industry is flourishing, there has been insufficient research exploring how, and to what extent, test preparation is effective. Existing literature seems to converge on the notion that test preparation, in most cases, has a positive effect on test performance (e.g., Appelrouth et al., 2017; Briggs, 2001; 2009; Bunting & Mooney, 2001; Hausknecht et al., 2007; Powers & Rock, 1999). However, the effect shown in those studies is inconsistent and inconclusive. Some studies detected a moderate to strong effect size (e.g., Hausknecht et al., 2007, $d = 0.70$), whereas others showed weaker or even negative effects (e.g., Briggs, 2009; McGaghie et al., 2010). In this case, an overarching quantitative synthesis of existing empirical research regarding the effect of test preparation, which would make it

possible to draw a general conclusion, is needed. Therefore, we present a meta-analysis to investigate the effect of test preparation activities in educational settings on test performance of large-scale tests, in comparison with no preparation or 'business-as-usual' conditions. This study also provides valuable insights into possible differential effects of test preparation by including specific characteristics of test preparation intervention design, strategy and material use, and the nature of target test in a moderator analysis.

2. Evaluating the evidence of test preparation

2.1 Working definition

Various terms have been used in the literature to describe teachers' and students' test preparation behaviours, such as coaching, teaching to the test, test-oriented instruction, test familiarisation, private tutoring, test-wiseness, etc. In this study, we use 'test preparation' as an holistic and inclusive term, which refers to 'any intervention procedure specifically undertaken to improve test scores, whether by improving the skills measured by the test or by improving the skills for taking the test, or both' (Messick, 1982, p.70). In this definition, all learning activities, no matter whether they were driven by teachers or students, as long as the main intention was to improve performance in a specific test, could be considered as test preparation activities. Therefore, test preparation is sometimes integrated within regular teaching and learning practices. However, not all learning strategies can be considered as test preparation strategies. Some general teaching and learning strategies that allow students to learn and understand new information or develop general cognitive and noncognitive abilities, such as collaborative learning and spaced learning, may be not directly related to test preparation. Based on the understanding of this definition, the indicator of the effects of test preparation activities could be either the observed gain in a test-taker's test scores after some form of test preparation has taken place, or the gap of test performance between those

who engaged in test preparation with their unprepared cohorts (Anastasi, 1981; Messick, 1982; Briggs, 2001).

A large number of empirical studies have reported teachers and students' engagement in test preparation activities (e.g., Anastasi, 1981; Appelrouth et al., 2015; Briggs, 2009; Lai & Waltman, 2008). In this article, we categorise test preparation activities as either student-driven or instructor-driven based on Briggs's (2009) classification. Student-driven test preparation has various forms, including but not limited to self-studying test preparation booklets, using online materials or computer programmes, practicing sample test items or repeatedly taking the tests, and asking experience from previous test-takers. Student-driven test preparation has been reported as the most used test preparation method (Briggs, 2009). In comparison, instructor-driven test preparation mainly comprises coaching or training courses on admission tests or other high-stakes tests in both formal schools and commercial educational institutes (Briggs, 2009; Bangert-Drowns et al., 1983).

2.2 Previous empirical research on the test preparation effect

Although the effect of test preparation has been investigated by a considerable number of empirical studies, only a few used a (quasi) experimental design. Most studies used self-report surveys to investigate participants' test preparation histories, then compared test performance of prepared students to those of unprepared students (e.g., Briggs, 2001; Briggs & Domingue, 2009; Powers & Rock, 1999; Hu & Trenkic, 2019). The effects reported were inconsistent. Briggs (2001) and Briggs and Domingue (2009) derived data from the National Educational Longitudinal Study in the U.S. to explore the relationship between students' use of different test preparation methods and the score gains on SAT or ACT. The results showed that the use of a private tutor and enrolment in commercial coaching courses were the only test preparation methods that had small but positive effects on SAT-Mathematics scores;

while no forms of test preparation had statistically significant effects on SAT-Verbal scores and ACT scores.

Studies conducted in different test contexts reported small or even negative effects of instructor-driven test preparation (i.e., Hu & Trenkic, 2019; Xie, 2013; Griffin et al., 2012). For example, Xie (2013) examined the test preparation effect on scores of College English Test, and the multiple regression and structural equation modelling results revealed that preparation for the test by drilling did have a significant effect on improving test scores (score gains 0.27 *SD*s), while other test preparation approaches such as developing general language skills did not affect test scores. Griffin et al. (2012) investigated whether attending commercial test preparation course affects performance in the Undergraduate Medical and Health Sciences Admission Test (UMAT). They found that enrolment in commercial coaching courses had inconsistent effects on students' performance in different sections of the test and for different levels of student ability. Their results showed that test preparation had no effect on student performance in the problem solving and 'understanding people' sections of the test, while it had positive effects for high ability students on their performance in non-verbal or abstract reasoning and negative effects for lower ability students.

Given the inconsistent findings of the aforementioned studies, it would be problematic to draw a conclusion about the test preparation effect. Besides, some studies using self-report surveys lacked control of confounding variables and ignored the heterogeneity of test preparation (Briggs, 2009). Thus, the effect of test preparation calculated by these studies is possibly inaccurate. Besides, specific characteristics of test preparation activities, such as material use, duration, and students' strategy use, etc., were rarely collected and reported in most self-report surveys. As a result, empirical research has so far been unable to provide a comprehensive picture of test preparation activities and their effects, thus a research synthesis is required.

2.3 Previous reviews

Tracing back to the 1980s and 1990s, a considerable number of review articles synthesising (quasi) experimental studies on test preparation effects were published primarily on coaching, but the effect reported in these articles varied. One of the earliest and influential meta-analytic reviews conducted by Bangert-Drowns et al. (1983), investigated the coaching effect on achievement tests using data from 30 controlled studies. This meta-analysis review reported a significant effect of coaching programmes on ACT scores (d = 0.25), which was affected by the level of training intervention. Specifically, they found that test preparation programmes designed to improve broad cognitive skills yielded the largest effects (d = 0.66) and extensive programmes which included drill practice had large effects (d = 0.43), while short test-oriented sessions had smaller effects (d = 0.17). Bangert-Drowns et al. argued that their finding was in line with what exam boards had long claimed - that long-term preparation on broad knowledge and abilities could have a larger effect than simple drill or practice on items (ETS, 1986). However, there were some inherent limitations to this research due to the imbalance in the number of studies included, raising concerns about the generalisability of the findings. Most interventions (22 out of 30) included in this meta-analysis were short test-oriented programmes, while only one included an intervention focused on broad skill training. The effect of coaching of broad skills was highly likely to be overestimated by pooling the effect size from only one study.

In the meta-analysis they published later on the effect of coaching on aptitude tests, Kuilk et al. (1984) found that aptitude tests such as SAT and intelligence tests were less affected by coaching than other aptitude tests. Coaching on the SAT had a small effect (d = 0.15), whereas coaching on other aptitude and intelligence tests had a substantial effect (d = 0.43). This finding was echoed by Becker's (1990) comprehensive synthesis of the effect of coaching programmes on SAT scores, in which she concluded that although test preparation

improved SAT scores in general, its effect was small and varied widely in terms of test domain. According to her study, test preparation interventions raised SAT-Verbal scores by 0.09 SDs and SAT-Math scores by 0.16 SDs, which implied that the math section in aptitude tests is more susceptible to test preparation than the verbal section.

Similar results were reported in a more recent systematic review of the effects of preparatory courses on SAT scores (Montgomer & Lilly, 2012). Ten experimental studies published between 1970-2000 related to the effect of SAT preparation courses yielded an average of 23.5 points of score gain on the Verbal subtest and 32.7 points gain on the Math subtest separately. Moreover, long coaching programmes (more than 8 hours) posted a significantly larger effect on SAT-Math scores than short programmes (8 hours or less).

In addition to reviews on coaching effects, there were also some reviews focused on the effect of taking tests repeatedly on score increases, which is known as the practice test effect or the retest effect. The earliest meta-analysis of practice effect was conducted by Kulik et al. (1984). Their study synthesised data from 40 studies related to the effect of taking practice forms of the test prior to the criterion aptitude or achievement test, revealed that taking practice tests that have identical form with the criterion test yielded a significantly larger effect on test scores (d = 0.42), compared with practice tests that used alternative forms (d = 0.23). Kulik et al. 's (1984) finding is broadly in line with two later meta-analyses on the retest effect for cognitive ability tests, conducted by Hausknecht et al. (2007) and Scharfen et al. (2018). Hausknecht et al. 's (2007) study reported an adjusted overall effect size of 0.26, while Scharfen et al.'s (2018) study revealed that retaking cognitive ability tests multiple times can lead to score gains of 0.5 SDs, and they also found that the retest effect stagnates after taking the test more than three times. However, considering that it is a gradual process for students to learn domain-specific knowledge and develop relevant abilities, whether

practice tests on their own can increase test scores on proficiency tests or knowledge tests is still questionable and needs further exploration.

To sum up, previous reviews identified a small coaching effect and practice effect on test performance. However, there were several limitations in these reviews, making it difficult to generalise the findings. To be specific, previous reviews on test preparation effects concentrated on the coaching effect for college admission tests in the U.S. context or the practice effect on cognitive ability tests, whereas the effect of test preparation on other proficiency tests or knowledge tests of a specific content domain in a wide range of educational testing contexts was not addressed by existing reviews, thus impeding our understanding of the variance of test preparation effect.

Moreover, most reviews were published decades ago and featured test preparation effect from articles published in 1960-1980. Findings drawn from these articles may not be applicable in today's educational environment. For these reasons, this meta-analysis is conducted to provide more up-to-date conclusions about test preparation effects and expand results to multiple educational contexts.

2.4 Theoretical underpinning of test preparation effect

To explain the effect of test preparation, various interpretations have been made in previous empirical studies and reviews, such as the improvement of memory and knowledge retention (e.g., Hausknecht, et al., 2007), the development of proficiency or abilities (e.g., Messick, 1982; Xie, 2013), the increase of test-wiseness (e.g., Madaus, 1998), the increase in motivation and confidence (e.g., Hong & Peng, 2007; Xie & Andrew, 2013) or the reduction of test anxiety (e.g., Messick & Jungeblut, 1981). Arendasy et al. (2016) summarised four possible models to explain the causes of test preparation effects from existing literature. The first postulates that score gains should be attributed to the familiarisation of test-taking and test structure, which reduces construct-irrelevant variance. For test-takers, the first few test

questions are easier because they do not need to familiarise themselves with the structure of the test. However, such gains caused by familiarisation cannot reveal students' abilities, because they are not relevant to the construct being assessed. The second model, according to Arendasy et al., states that engaging in test preparation improves test-wiseness and test-taking skills, such as guessing and eliminating wrong answers, which can improve the likelihood to correctly solve test items, without affecting the underlying construct. The third model ascribes the score gain to teaching construct-relevant materials, which further results in the development of test-specific cognitive abilities or an increase in domain-specific knowledge. The fourth model is based upon the premise that the increase in test scores is caused by the improvement of generalisable cognitive abilities.

These four models form a comprehensive theoretical framework of test preparation effect. There are numerous potential factors influencing the effect of test preparation that can be deduced from the four models, which are investigated by the current paper. Moreover, given that Arendasy et al 's (2016) explanations were mainly drawn from retest effect literature, we will examine the applicability of this framework in the test preparation context and explore other possible underlying mechanisms of test preparation effect.

2.5 Potential moderator variables

In the aforementioned empirical studies and reviews, it has been shown that instructional characteristics in test preparation intervention (e.g., methods, the use of practice tests, instructional emphasis, duration, transferability) contribute to the effect on test score improvement. In addition, a body of literature has indicated that test-specific features (e.g., domain, type of tasks, cognitive ability) and individual differences (e.g., age, ability level, prior attainment, previous test experience) explain variations in test preparation effects (e.g., Arendasy, et al., 2016; Briggs, 2009; Kulik, et al., 1984; Powers, 1999; Powers & Swinton, 1984). In line with previous empirical evidence, we examined whether, and the extent to

which, different implementations of test preparation in different testing contexts have varied effects on individuals' test performance. Moreover, the effectiveness of test preparation is likely to be contingent on study design (Becker, 1990; Briggs, 2009; Hausknecht, et al.,2007), thus design characteristics were also examined in our moderator analysis.

2.5.1 Design characteristics of test preparation intervention

The following moderator variables have been selected as test preparation design characteristics: type of intervention, level of intervention, instruction on strategy use, material use, and duration of the intervention.

Type of intervention refers to the specific test preparation methods, which were organised into three basic categories, including school courses, commercial courses, and self-study relevant materials. Briggs (2001) reported that commercial courses have larger effects than other methods. Previous reviews mainly examined the effect of commercial courses, whereas the effect variance related to different test preparation methods has not been explored in previous meta-analyses.

We also analysed the *breadth of intervention,* which comprised three levels based on the instructional and practical emphasis: 1) broad on a general knowledge/skill area or test-associated curriculum, without a narrow focus on the test; 2) specific instruction on test-specific content; and 3) narrow (or drilling) instruction on exactly the same tested items. The level of instruction might explain the variance of the effects between the studies. Kulik et al. (1983) found in their meta-analysis that coaching programmes designed to improve broad cognitive abilities yielded the largest effect. In contrast, some empirical studies showed that specific test preparation focusing on test-related content or drilling on tested items has significant effects on test scores, although the predictive validity was not improved in this way (Anastasi, 1981; Allalouf & Ben-shakhar, 1998; Arendasy, et al., 2016; Xie, 2013; Hu & Trenkic, 2019).

Moreover, we examined whether the *instruction on strategy use* is a potential moderator variable. It has been found that the strategies students' use during test preparation are linked to test performance (Purpura, 1999; Phakiti, 2003). Possible effective strategies include test preparation management, test-taking skills, memorisation, socio-affective strategies, and broad knowledge/skills development strategies (Hsiao & Oxford, 2002; Oxford, 1990; Xie, 2013). Xie (2013) examined the effectiveness of test preparation strategies, finding that drilling and test taking skills were the most effective ways affecting test scores, although the effect was small.

Furthermore, we investigated if different *material use* in interventions could lead to the variance in test preparation effect. The effect of varied use of materials has rarely been examined by previous meta-analyses, and we expect to identify the kinds of materials that could improve test performance effectively. The use of practice test, sample items, workbooks and computer programmes were investigated to see if the test preparation effect would be moderated by these materials. Finally, we included *test preparation time* of the intervention to see if the amount of contact hours and overall duration of the intervention (weeks) might explain the variance of the effects between studies. Bangert-Drowns et al. (1983) found that the duration of coaching programme positively correlated with the magnitude of the coaching effect.

2.5.2 Test-specific characteristics

The following two moderating variables associated with test specific features were seen as relevant: academic domain and nature of the test.

We examined whether the *academic domain* assessed by the test moderated the test preparation effect. Research indicated that tests consisting of analytical or quantitative tasks were more coachable and benefited more from repeated test-taking than did verbal tasks

(Hausknecht, et al., 2007; Kulik, et al., 1984). We coded the subject domain to investigate variance of test preparation effects were affected by domain type.

Moreover, we included *nature of the test* in the moderator analysis to see whether test preparation was as effective for high-stakes tests as it was for low-stakes tests; for large-scale tests as it was for course tests; and for curriculum-based tests as it was for non-curriculum-based tests. Previous reviews mainly focused on the test preparation effect on high-stakes tests, especially college admission tests. Only very limited number of studies examined test preparation effects on low-stakes tests or course tests (e.g., Burnner et al, 2007).

2.5.3 Student characteristics

There is plentiful evidence that individual differences are possible moderators of students' test preparation effect (Appelrouth, Zabrucky, & Moore, 2015; Xie & Andrew, 2013). For example, Kulik et al. (1984) found that the average test preparation effect is highest for high-ability students (ES=0.82). Similarly, Griffin et al. (2012) found that coaching had a significant effect on the reasoning section of UMAT for high ability students, while for those of lower ability, the effect was negative.

Besides, it has been indicated that students' past test experience might moderate test preparation effect. Briggs and Domingue's (2009) study revealed that only students with previous test-taking experience can significantly benefit from private tutoring on ACT.

In addition, test preparation processes could also operate differently for students from different education stages and age groups, due to differences in self-regulated learning abilities which tend to increase at higher ages. Kitsantas (2002) argued that students who use more self-regulatory skills (such as goal setting, keeping records and monitoring, and self-evaluation) during preparation have higher test performance.

We therefore investigated whether student characteristics such as *education level* and *previous experience of test-taking* influence the degree to which test preparation can be used to effectively improve test performance and long-term development.

2.5.4 Study design

It is also possible that studies with a more rigorous experimental study design yielded greater effects than studies with less quality. Previous review articles indicated that randomised experiments reported significantly higher effects of test preparation on test success than quasi-experiments (Kulik, et al., 1983; 1984), however, only a small proportion of studies included in previous reviews used a randomised design with control group (Bangert-Drowns et al., 1983; Briggs, 2009). Thus, in this study, we also examined whether *Allocation type* and *Control group type* are possible moderators that predicted the magnitude and direction of test preparation effect. To note that control conditions in most studies are some forms of 'business as usual' (e.g., Alderman & Powers, 1980; Bookman, 1981), while some studies compare two types of test preparation practice, for example, intensive test preparation vs. instructions on general skills or broad knowledge (e.g., Ercanbrack, 1999; Nishitani, 2006).

3. Method

3.1 Literature search

The literature search was initially carried out between December 2019 to May 2020 using five electronic bibliographic databases: *ERIC, ProQuest*, *PsycInfo, Web of Science* and *GoogleScholar*. In June 2021, the first author checked these databases again to locate newly published relevant studies. To access the initial body of relevant literature, the publication date was not specified. Key- and subject- word searches were conducted to capture all possible literature. The standard truncation (*) was also used for retrieving variations of the

search terms. Searches utilized the following words and word combinations: all (test prep* OR test preparation OR shadow education* OR coaching OR private tutor OR tutor*) AND all (test* OR test-tak*) AND all (performance OR test score* or achievement* OR attainment*) OR all (effect OR washback* OR practice effect*). In addition, a secondary literature search was conducted by scanning reference lists from recent meta-analytic investigations of test preparation effects and practice effects on educational tests (i.e., Adesope et al. 2017; Hauskenecht, et al. 2007). The Social Science Citation Index database was also used to search journal articles that referenced these meta-analysis articles. Grey literature that may be applicable to this study such as dissertations, conference papers and reports were also included. The above electronic and manual literature covered all of the papers that were available through 2021.

3.2 Inclusion and exclusion criteria

Of the 4277 retrieved records in total, 1309 duplicates were removed. We then screened titles, abstracts, keywords, and subheadings of each article for possible inclusion. For these studies in which the abstract did not provide sufficient information, we read sections related to the research design to gather more information. Detailed inclusion and exclusion criteria are listed in Table 5. Studies not examining issues related to test preparation effect or not using quantitative methods were excluded. Besides, we identified substantive studies examining the retention of using specific learning activities (e.g., retrieval practices or spaced learning) on memory-based tasks in classroom settings (e.g., McDaniel et al., 2015; Roediger et al., 2011). Most of them were not included in this meta-analysis, mainly because the intervention in these studies were not designed to help students to prepare for a specific test, and large-scale educational tests were not used as the measure.

After the screening stage, 134 studies were retained. We read the full texts of all 134 studies to determine whether they meet all inclusion criteria. The literature searching and processing stages are showed in Figure 8.

Table 5 Inclusion and exclusion criteria.

Inclusion	**Exclusion**
1. It examined the effect of test preparation activities as listed in the working definition. The treatment condition(s) had to include any form of test preparation intervention (i.e., test preparation courses or computer programmes) or instructional intervention towards test-taking for one specific test.	Articles focused on other educational activities which main purpose was not for test preparation, even they used test scores to indicate students' attainment.
2. It investigated test performance of large-scale tests designed for educational purposes and conducted in real life educational contexts.	Articles investigated preparation effect on course tests, other small-scale tests, such as memorisation tests developed by researchers, or other tests not used in educational contexts such as cognitive tests in operational contexts, and neuropsychological tests administered for clinical diagnoses.
3. It examined test preparation effect on a non-self-reported measure of test performance.	Articles used self-reported test scores as dependent variables.
4. It included either an experimental or quasi-experimental design with at least one treatment group and one control group, or report both pre-test and post-test scores in the case of within-subject design, which allowed to draw causal inferences regarding the effect of test preparation on test performance,	Articles that reported effect that cannot be attributed to test preparation intervention were excluded. For example, some articles asked students to report whether they participated in test preparation or not, and made comparisons among their scores.
5. Sufficient data (i.e., means, standard deviations, sample sizes) should be reported	

	in each included study to enable the computation of effect sizes.	
6.	Included studies could be conducted at any education level and were not limited to any specific study site.	
7.	Only SAT-related studies conducted after 1978 and ACT-related studies conducted after 1985 were included.	Any SAT-related studies conducted before 1978 and ACT-related studies conducted before 1985 were excluded, as sample tests of SAT and ACT were not provided by the exam bodies to the public before the two years (A (Mostly) Brief History Of The SAT And ACT Tests, n.d.).
8.	Studies published in English were included to ensure the reliability and coherence of coding the key elements of test preparation intervention.	

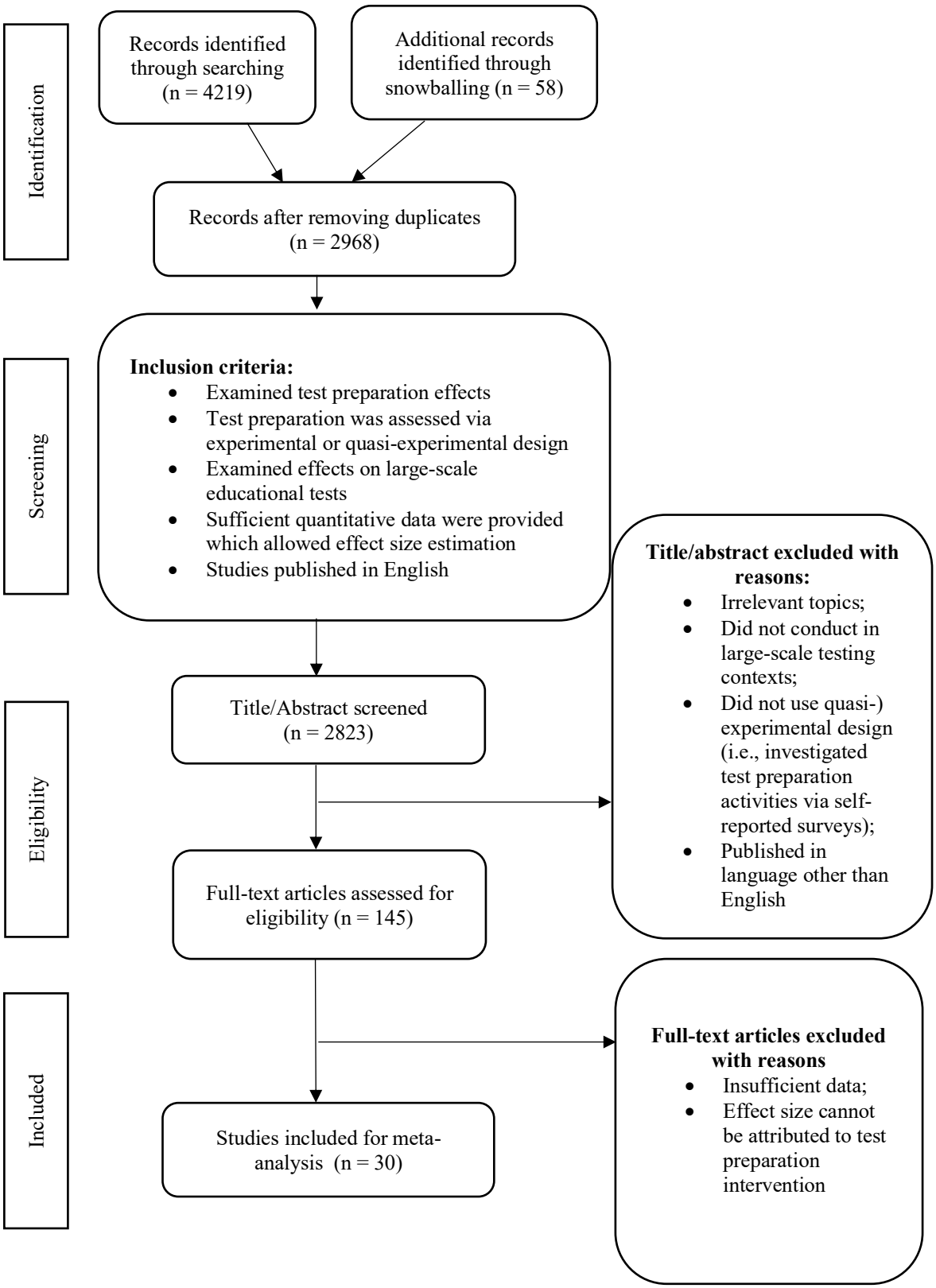


Figure 8 Flow diagram of literature search and processing of records.

3.3 Data extraction

The included studies were systematically analysed by making use of two coding schemes: one on the study level and another scheme on the effect size level. In the first scheme, five groups of variables were coded for each study, including general study descriptors (e.g., study type and publication year), test preparation characteristics (e.g., duration of intervention, type of intervention, level of instruction), student characteristics (e.g., gender, ability level, education level), test-related variables (e.g., subject domain, nature of test) and study characteristics (e.g., type of comparison group, allocation). In the second coding scheme, raw data were collected to calculate the particular effect size, and all information about the dependent variable (the test scores, or assessment on long-term retention) were gathered.

We took an iterative approach to extract data whereby coders refined the classification of each variable as they progressed through the included studies to ensure that the classifications best characterised the literature. The utilised classification of each coding variable is listed below.

Nature of Publication

Publications were classified as either: journal articles, dissertations, conference papers, exam board reports, industry reports, and others. The year of each publication was specified.

Country

The country in which the study conducted was recorded and classified as either: Mixed, US and others. The name of the country was also specified.

Nature of test

The nature of the test was classified as either curriculum-based and non-curriculum based test. Taking the two widely used college admission tests in the US as examples, ACT is a curriculum-based test while SAT is a non-curriculum based test.

Stakes of test

The stakes of the test which students prepared for was classified as either high stakes or low stakes. High-stakes tests refer to tests with important consequences for students (Madaus, 1988), while low-stakes tests do not have important consequences associated with test performance. It is important to note that whether the test is high-stakes or low-stakes is relative to test-takers. For example, in one of the included studies, Hardison and Sackett (2008) examined the effect of short-term rule-based coaching on standardised writing tests, using writing prompts from College Level Examination Program (CLEP) to examine students' writing performance. The CLEP can be high-stakes for students who want to use it to receive college credit. However, in Hardison and Sackett's (2008) study, students were not attending the real CLEP but used CLEP prompts to assess their writing improvement after the intervention, thus the test in this study was coded as low-stakes.

Test type

Based on the evaluation of all included articles, we classify the large-scale educational tests in the meta-analysis as either: college admission test (such as SAT, ACT), English language test (such as TOEFL, IELTS), and others.

Subject domain of the test

The subject domain focused by the target test in each individual study was classified as either: verbal, quantitative/mathematics, science, foreign language, and mixed. If outcomes from multiple domains were reported in the article, then the outcome on each domain were recorded and separate effect sizes were reported. If the study reported an overall combined score of multiple domains, then the subject was coded as 'mixed'.

Format of the test

The inclusion of multiple-choice questions in the test was classified as yes or no, as was the inclusion of short-response questions. In some included articles, authors did not mention

information about the test format, thus the coders checked the official website of the test to gather the information.

Test administration

The administration of the test was classified as either mock, indicating that mock tests were used for the post-test; or real, indicating that participants in the study sat the real test after test preparation intervention.

Type of test preparation intervention

The test preparation intervention was classified as commercial courses, in-school courses, after-school courses or self-study. To note that except for specialised courses provided by commercial test preparation centers, we regarded paid courses offered in public or private schools as commercial courses as well, because those paid courses were selective and some were taught by teachers from testing centers or test preparation institutes (e.g., Filizola, 2008). Courses provided by schools or universities were coded as either in-school courses or after-school courses (i.e., extra-curricular activities).

Breadth of test preparation intervention

We classified the breadth of test preparation intervention as broad, specific, or narrow. When the intervention focused on a general knowledge/skill area (but may have some overlap subjects with the tested content), it was coded as broad. When the intervention targeted at the same subject or area with the test, but did not drill on the test, it was coded as specific. When the intervention only focused on the test-specific content and format or drilled on the exact same tested items, it was coded as narrow.

Strategy use

The provision of inclusion in each of five test preparation strategies was classified as yes or no. The five strategies investigated by the current meta-analysis included:

Test preparation management (TPM) strategy, which resembles metacognitive strategies and refers to test preparation practices through analysing test papers to identify frequently assessed areas, learning marking rubrics and sample responses to evaluate own answers, conducting self-assessment of personal strengths and weaknesses, and effectively managing time for own test preparation process;

Test Taking Strategy (TTS), which refers to the practices that test takers explore and practice test-taking skills for different sections of the test, including guessing, context clues, eliminate incorrect answers, take notes in the margin, and time management for test-taking, etc;

Drill, which refers to the intensive and repetitive practice of a narrow range of skills and knowledge tested by the test, such as memorising frequently tested knowledge or facts, and repeatedly practicing past test questions and remembering exemplar responses;

Socio-affective (SOAF) strategy, which refers to test-takers' use of social strategies to seek support from teachers and peers, and their use of affective strategies to motivate themselves and to reduce test anxiety;

Broad Knowledge Skills Development (BKSD) strategy, which refers to the learning strategies that test takers use to develop broad knowledge or skills via extensive exposure to and functional uses of knowledge or skills in authentic contexts.

If relevant features of any of the above strategies were described in a study, then the study was coded as 'yes' for the use the particular strategy. However, if no relevant feature was mentioned in any of the above strategies, then the strategy was deemed not to have been used in the study and was coded as 'no'.

Material use

The use of each of the four types of materials was classified as yes or no. Materials included *commercial test preparation workbooks*, *timed practice tests*, *sample items*, and

computer-based programmes. If an included study did not mention the use of a specific material, we assumed it did not use the material and coded it as 'no'.

Time and duration of the intervention

We recorded the time (indicated by number of contact hours) and duration (indicated by weeks) of each intervention.

Education level

Participants' education level was classified as either: tertiary, secondary, primary or mixed. If the study invited participants from a wide population and mixed with their educational levels, it was coded as mixed (see Fransworth, 2013 as an example).

Previous test experience

We recorded participants' previous test experience. Four categories of participants' experience were found in the included studies: yes, no, parts of students have relevant experience, or not mention.

Allocation

The ways how participants were allocated to conditions were classified as random allocation or non-random.

Comparison group

We identified three types of comparison groups in the included studies: no test preparation condition, different test preparation conditions, and using pre-test as the control (within participant design). Participants in a no test preparation condition usually did not receive any instruction or received instruction after the intervention (e.g., Moss 2012). Some included studies compared the effects of different types of test preparation, which could be characterised as a business-as-usual condition in many instances; that is, two versions of a course were run – one provided specific test preparation tasks or instructions, and one provided the normal course as usual. For example, Winke and Lim (2017) compared the

effects of three different interventions: explicit instruction, implicit instruction, and control, on second language listening test performance. We therefore extracted two comparisons from this study: explicit instruction vs. control and implicit instruction vs. control. The control condition provided students with instructions on American culture, which was regarded as broad instruction on second language learning, thereby was coded as different test preparation condition. In addition, some studies examined the effects of different test preparation interventions and provided detailed description on each intervention (e.g., Pan, 2016) compared effects of traditional test preparation practices and non-traditional test preparation practices). In this case, rather than comparing the two different test preparation methods directly, we regarded each test preparation intervention as a within subject design and used the pre-test results as the control. In this way, we were able to fully capture the features of each intervention and explore how these features influenced test preparation effects.

3.4 Coding and inter-rater reliability

Three rounds of coding training were performed to allow coders to discuss and refine the classification to well characterise each individual article. Within the coder training, 20% of included studies ($N = 6$) were coded by all three coders (with an interrater agreement above .80 for each article). Disagreements were discussed and resolved, while the coding scheme was modified based on the discussion. After coding trials, the remaining articles were randomly grouped into two sets. The first author coded all included articles independently, and the other two coders coded 50% each, with random allocation. The initial interrater reliability was 89%. The discrepancies were resolved through discussion.

3.5 Statistical analytical strategy

A random-effects meta-analysis was conducted using R version 3.6.3 (R Core Team, 2019). For the analyses, we followed the procedure described by Borenstein et al. (2009) for effect size calculation, integration, and meta-regression for moderator analysis. We used the standardised mean difference effect size to measure the effects of test preparation interventions between treatment and control groups on student test performance. The data on the effects of included studies was gathered in an Excel sheet. Cohen's *d*, variance, and *SE* of Cohen's d were computed in Excel. Then, Hedge's *g* was calculated and effect aggregation was performed in R Studio using the 'metafor' and 'robumeta' packages. When both pre-test and post-test data were available, pre-test data were used to adjust the effect sizes by subtracting the pre-test effect sizes from the post-test effect sizes and adding up the variance of both effects.

To note that many studies included in our meta-analysis either reported multiple outcomes using the same participant groups or nested several treatments and control conditions in one study. Including those dependent effect sizes can bias the results of the analysis, as studies reported more effect sizes would take larger account. Therefore, we employed robust variance estimation (RVE) to handle dependent effect sizes (Tanner-Smith, et al. 2016), which accounts for both hierarchical effect cases (i.e., multiple experiments are nested) and correlated effect cases (i.e., different measures with same participants) with small-sample corrections.

Heterogeneity in the effect sizes was estimated using the Cochran Q-test. The I^2 statistic indicates that the proportion of the variation in effect sizes is due to true between-study heterogeneity rather than sampling error, and $I^2 > 75\%$ represents considerable heterogeneity (Higgins, et al., 2003). Meta-regressions and subgroup analysis were performed to examine the moderating effects of various factors that differed across studies.

To investigate the moderating effects of possible factors that varied across studies, we conducted both sub-group analyses and meta-regressions. Although subgroup analysis could recognise whether there were differences between subgroups by determining whether the confidence intervals around their effect sizes overlap, heteroscedasticity or multicollinearity may lead to biased estimates (Steel & Kammeyer-Mueller, 2002). To examine the overall impact of a moderator, we performed separate meta-regressions for each predictor.

3.6 Assessment of publication bias

Publication bias is one of the threats to the validity of meta-analyses (Van Aert et al. 2019). It occurs as a result of some studies not being published due to the lack of statistical significance in their results (Banks et al., 2012). Carter et al. (2017) suggested if the probability of publication bias is high, the analysis should not rely only on the random effect model. For assessing and reporting publication bias, we conducted a moderator test between published and unpublished studies to explore the possibility of published studies had significantly higher effect sizes than unpublished studies. Funnel plots were also produced in R Studio to inspect asymmetry and heterogeneity. Considering that we have a relatively small sample of empirical studies, the funnel plots and Egger's regression test were used to evaluate whether small-study effects existed.

4. Results

4.1 Descriptive statistics

Overall, findings from 30 primary studies reporting 97 effect sizes were extracted for the meta-analysis. Descriptive information of all included studies is reported in Table 6. The majority of studies investigated test preparation effects on college admission tests such as ACT and SAT ($N = 13$) and English proficiency tests such as IELTS ($N = 12$); while only a small number of studies ($N = 5$) were conducted in other testing contexts. Most studies were

conducted in the U.S. (N = 17) while seven studies were conducted in Asian countries and regions. The majority of interventions reported by the studies included (N = 28) were led by instructors, while only two studies investigated students' self-preparation using computer programmes and handouts prepared by teacher. Most of the studies reported interventions that prepared students for one specific test and used the criterion test to measure students' performance. However, four studies utilised different tests to examine the transferability of test preparation effects.

Table 6 Descriptive characteristics of the included studies.

Authors	Year	Type	Region	Test	Domain	Ed. Level
Alderman & Powers	1980	Journal article	US	SAT	Verbal	Secondary
Allalouf & Shakhar	1988	Journal article	Israel	PET	Mixed	Secondary
Bailey & Judd	2018	Journal article	South Korea	TOEIC	English	Tertiary
Bookman	1981	Dissertation	US	Comprehensive Test of Basic Skills	Mathematics	Secondary
Brunner, et al.	2007	Journal article	German	PISA	Reading & Mathematics	Secondary
Bunting & Mooney	2001	Journal article	UK	11-Plus	Mixed	Primary

Farnsworth	2013	Journal article	US	Oral English Tests	English	Mixed
Filizola	2008	Dissertation	US	SAT	Mixed	Secondary
Gaye	2001	Dissertation	US	ACT	Mixed	Secondary
Green	2007	Journal article	UK	IELTS	English	Tertiary
Hardison & Sackett	2008	Journal article	US	Writing Tests	Writing	Tertiary
Holmes & Keffer	1995	Journal article	US	SAT	Verbal	Secondary
Holthaus	2008	Dissertation	US	SAT	Mathematics	Secondary
Justus	2010	Dissertation	US	ACT	Mixed	Secondary
Khodabakhshzadeh	2017	Journal article	Iran	IELTS	English	Tertiary
Lane, et al.	2009	Journal article	US	ACT	Mixed	Secondary
Lee	2019	Journal article	Taiwan	TOEIC	English	Tertiary
McMann	1994	Dissertation	US	ACT	Mathematics	Secondary
Mitchell, et al.	2016	Journal article	US	DEA	Reading	Primary
Moss, et al.	2012	Exam board Report	US	ACT	Mixed	Secondary
Nishitani	2006	Book chapter	Japan	TOEIC	English	Tertiary
Pan	2016	Journal Article	Taiwan	GEPT	English	Tertiary

Parrott	2012	Dissertation	US	ACT	Mixed	Secondary
Rao	2003	Exam board report	Fiji	IELTS	Tertiary	Tertiary
Robb & Ercanbrack	1999	Report	Japan	TOEIC	English	Tertiary
Roberts & Oppenheim	1966	Dissertation	US	PSAT	Mixed	Secondary
Takallou, et al.	2015	Journal article	Iran	Konkoor	English	Secondary
Trenkic & Hu	2021	Journal Article	Mainland China	IELTS	English	Tertiary
Winke & Lim	2017	Journal article	US	IELTS	English	Tertiary
Zuman	1988	Journal article	US	SAT	Mixed	Secondary

4.2 Overall meta-analysis of the effect of test preparation

In the overall model, a significant positive effect of test preparation on criterion test performance was found: $g = 0.38$, $se = 0.06$, 95%$CI = 0.26\text{-}0.50$, $p < .001$. A forest plot of all recorded effect sizes is provided in Figure 9. A sensitivity analysis indicated that the effect size estimates ranged from 0.372 to 0.384 with different values of rho, which was small to medium.

The heterogeneity between the studies' effect size was significant with $I^2 = 81.1\%$, $Q(97) = 508.40$, $p < .001$, which shows that the test preparation interventions do not share the same true effect size and the heterogeneity is high (Higgins & Thompson, 2002). The variance of the true effect sizes is estimated at $\tau^2 = 0.11$. This supports the use of a meta-

regression and sub-group analysis in order to explain the observed heterogeneity in effect sizes.

This meta-analysis also investigated the transferability of the test preparation effect. A small number of included studies (N = 4, k = 5) explored the transferability of preparing for a specific large-scale test to other tests within the similar domain, such as whether the IELTS preparation course could improve the performance of the Online Oxford Placement Test (Trenkic & Hu, 2021), or if the training on SAT writing skills benefited other writing tasks such as writing personal statements or complaint letters (Hardison & Sackett, 2008). The results showed that the effect of test preparation seems to be unidirectional to a specific test and unable to be transferred, as the effect size was small and not significant: g = 0.06, se = 0.08, p = 0.502. However, we should be very careful with the interpretation of this result, as our data were drawn from a limited number of studies. Further research is needed to justify the transferability of the test preparation effect.

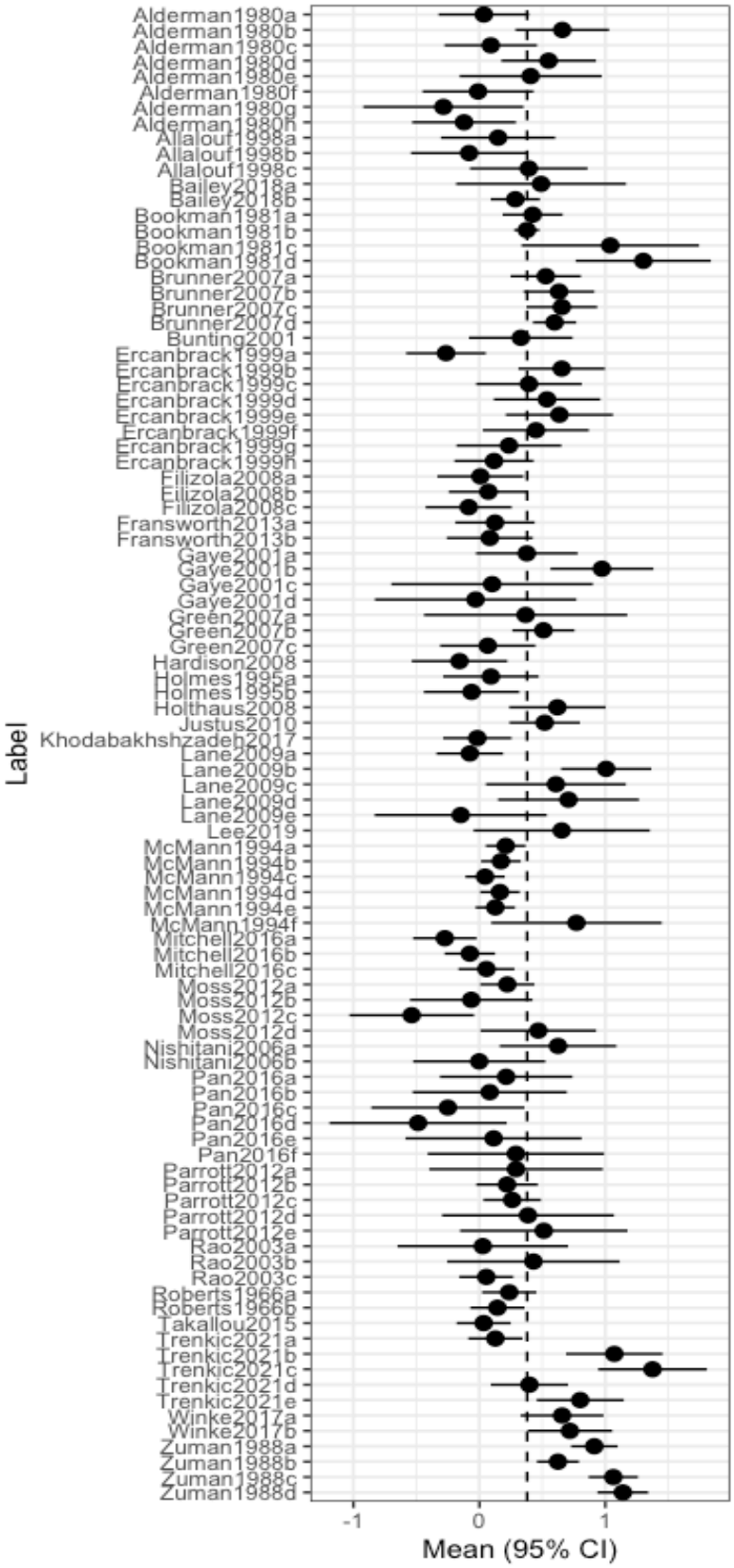


Figure 9 Forest plot of all recorded effect sizes.

4.3 Moderator analysis

We examined whether test specific characteristics, test preparation related factors, student-related factors and study design characters explained the variance in effect sizes. Effect sizes for sub-groups are presented in Table 7. Results of the meta-regressions are presented in Table 8.

Table 7 Results of the sub-group analysis.

	N	k	Sample size	g	SE	I^2	*p*
Nature of the test							
Test Stakes							
High-stakes tests	20	71	9183	0.42	0.08	82.73%	<.001
Low-stakes tests	11	26	3083	0.27	0.08	66.23%	0.06
Relationship with Curriculum							
Curriculum-based tests	18	56	4020	0.36	0.08	62.8%	<.001
Non-curriculum-based tests	12	41	8246	0.29	0.09	87.22%	0.008
Type of tests							
College admission tests	14	49	6262	0.35	0.08	81.44%	<.001
English proficiency tests	11	35	2895	0.47	0.11	77.80%	0.002
Other tests	5	13	3109	0.28	0.11	85.85%	0.073
Subject Domain							
Verbal	13	29	3999	0.28	0.09	85.79%	0.011
Mathematics	13	23	2989	0.31	0.09	75.74%	0.003
EFL	13	37	3291	0.52	0.10	80.82%	<.001
Others	4	4	702	0.33	0.28	96.08%	0.327
Mixed	4	4	1285	0.37	0.14	87.64%	0.083
Test Format							
MCQ questions Included							
Yes	25	82	10900	0.4	0.07	82.25%	<.001
No	7	15	1366	0.29	0.04	49.28%	<.001
SR questions Included							
Yes	4	10	1147	0.34	0.14	67.13%	0.103
No	28	87	11119	0.35	0.06	79.05%	<.001

Test Administration							
Real test conditions	13	43	5493	0.40	0.07	82.25%	0.002
Mocks	18	54	6773	0.29	0.04	49.28%	<.001
Type of Test Preparation Intervention							
After-School courses	6	25	3434	0.38	0.16	90.85%	0.072
In-School courses	15	45	5742	0.36	0.09	82.06%	<.001
Commercial courses	6	20	2110	0.38	0.12	58.19%	0.030
Self-study	2	3	112	0.08	0.35	54.78%	0.851
Unknown	3	4	868	0.36	0.12	15.95%	0.153
Breadth of Test Preparation Intervention							
Broad	7	12	1551	0.13	0.08	74.79%	0.166
Specific	13	48	7359	0.38	0.08	80.59%	0.001
Narrow	14	37	3356	0.36	0.08	60.6%	<.001
Strategy Use for Test Preparation							
Test preparation management							
Yes	20	45	4938	0.46	0.08	80.3%	<.001
No	15	52	3083	0.27	0.07	73.34%	0.002
Test-taking skills							
Yes	24	75	9301	0.43	0.07	81.57%	<.001
No	9	22	2965	0.16	0.06	55.22%	0.043
Memorisation							
Yes	23	73	9224	0.31	0.06	66.86%	<.001
No	10	24	3042	0.47	0.11	84.08%	0.002
Socio-affective strategies							
Yes	7	16	1790	0.53	0.14	84.92%	0.009
No	26	81	10476	0.34	0.07	76.02%	<.001
Broad knowledge/skill development							
Yes	20	56	5933	0.37	0.08	74.39%	<.001
No	16	41	6333	0.36	0.08	83.02%	<.001
Material Use for Test Preparation							
Practice tests							
Yes	9	23	2521	0.31	0.09	66.97%	0.012
No	22	74	9745	0.40	0.07	82.67%	<.001
Sample items							
Yes	26	75	9457	0.38	0.06	74.24%	<.001

No	6	22	2809	0.30	0.16	89.95%	0.114
Commercial Workbook							
Yes	13	49	6626	0.47	0.1	82.78%	0.001
No	19	48	5640	0.23	0.05	65.11%	<.001
Computer programme							
Yes	2	3	12172	0.41	0.04	0%	0.061
No	29	94	94	0.37	0.06	81.03%	<.001
Time Use for Test Preparation							
Contact hours							
Less than 10 hours	14	38	6243	0.28	0.09	83.03%	0.008
Between 10-30 hours	12	30	3506	0.29	0.07	56.34%	0.002
More than 40 hours	10	29	2517	0.41	0.11	75.32%	0.005
Duration							
Within 4 weeks	12	33	4178	0.28	0.08	62.55%	0.004
Between 4 weeks to 12 weeks	11	28	4238	0.32	0.09	83.75%	0.006
More than 12 weeks	12	36	3850	0.38	0.09	73.24%	0.003
Students' Education Level							
Primary	1	1	552	0.33	0.07	82.88%	<.001
Secondary	18	61	8633	0.33	0.07	82.88%	<.001
Tertiary	10	31	2261	0.50	0.11	78.54%	0.002
Mixed	2	5	726	0.10	0.13	23.57%	0.569
School type							
Private	3	8	383	0.39	0.16	12.93%	0.143
Public	11	40	6708	0.31	0.10	88.82%	0.012
Students' Previous Test or Test Preparation Experience							
Yes	4	12	1404	0.69	0.13	82.88%	0.013
No	11	37	4572	0.25	0.06	66.46%	0.003
Parts of students have relevant experience	6	23	2498	0.29	0.11	54.91%	0.06
Not mention	9	25	3698	0.44	0.14	80.71%	0.013
Random Allocation							
Yes	13	35	3991	0.49	0.08	62.03%	<.001
No	17	62	8181	0.30	0.08	84.32%	0.002
Participant Design							
Between	24	76	10215	0.34	0.06	71.22%	<.001

Within	6	21	1957	0.51	0.14	91.95%	0.015
Comparison							
No test preparation condition	19	59	8466	0.33	0.07	72.09%	<.001
Different test preparation conditions	8	20	1840	0.34	0.13	75.33%	0.033
Pre-test	4	18	1866	0.56	0.19	94.09%	0.062

Table 8 Results of meta-regressions.

	b	**SE**	**CI Low**	**CI Upper**	***p***
Nature of the test					
Test Stakes					
Intercept	0.43	0.08	0.27	0.59	*<.001*
Low-stakes tests	-0.15	0.11	-0.38	0.08	*0.194*
Curriculum					
Intercept	0.35	0.07	0.21	0.49	*<.001*
None-curriculum-based tests	-0.07	0.11	-0.28	0.14	*0.502*
Type of tests					
Intercept	0.35	0.08	0.17	0.54	*<.001*
English proficiency tests	0.12	0.14	-0.18	0.41	*0.420*
Other tests	-0.08	0.14	-0.41	0.25	*0.610*
Subject Domain					
Intercept	0.49	0.10	0.27	0.70	*<.001*
Verbal	-0.25	0.11	-0.48	-0.03	*0.031*
Mathematics	-0.17	0.13	-0.43	0.10	*0.210*
Others	-0.12	0.32	-1.05	0.81	*0.732*
Mixed	0.01	0.15	-0.51	0.53	*0.942*
Test format					
Included MCQ questions					
Intercept	0.29	0.03	0.20	0.38	*<.001*
Yes	0.11	0.08	-0.08	0.30	*0.202*
Included SR questions					
Intercept	0.33	0.05	0.23	0.44	*<.001*
Yes	-0.16	0.08	-0.32	0.01	*0.060*
Test administration					

Intercept	0.35	0.07	0.21	0.50	*<.001*
Real test conditions	0.06	0.12	-0.20	0.32	*0.636*
Type of Test Preparation Intervention					
Intercept	0.39	0.17	-0.08	0.85	0.083
In-school courses	0.01	0.20	-0.49	0.46	0.954
Commercial	-0.01	0.21	-0.49	0.46	0.948
Self-study	-0.148	0.38	-2.70	2.33	0.691
Unknown	-0.03	0.20	-0.59	0.53	0.888
Level of Test Preparation Intervention					
Intercept	0.18	0.07	0.04	0.33	0.011
Specific	0.19	0.08	0.04	0.34	0.014
Narrow	0.14	0.07	0.01	0.29	0.048
Strategy Use for Test Preparation					
Test preparation management					
Intercept	0.23	0.06	0.10	0.37	*0.003*
Yes	0.25	0.10	0.04	0.45	*0.021*
Test-taking skills					
Intercept	0.18	0.06	0.04	0.33	*0.022*
Yes	0.26	0.09	0.05	0.46	*0.019*
Memorisation					
Intercept	0.51	0.11	0.24	0.77	*0.002*
Yes	-0.19	0.13	-0.47	0.09	*0.175*
Socio-affective strategies					
Intercept	0.35	0.06	0.21	0.48	*<.001*
Yes	0.18	0.16	-0.20	0.55	*0.311*
Broad knowledge/skill development					
Intercept	0.38	0.09	0.19	0.57	*0.001*
Yes	0.00	0.11	-0.23	0.23	*0.971*
Material Use for Test Preparation					
Workbook					
Intercept	0.27	0.07	0.11	0.42	*0.002*
Yes	0.24	0.11	0.00	0.48	*0.046*
Practice tests					
Intercept	0.41	0.07	0.25	0.56	*<.001*
Yes	-0.1	0.12	-0.35	0.15	*0.402*

Sample items					
Intercept	0.38	0.19	-0.14	0.90	*0.112*
Yes	0.00	0.20	-0.49	0.49	*0.998*
Time Use for Test Preparation					
Contact hours					
Intercept	0.35	0.07	0.21	0.50	*<.001*
	0.00	0.00	-0.01	0.01	*0.644*
Duration					
Intercept	0.35	0.08	0.17	0.53	*0.001*
	0.00	0.01	-0.02	0.02	*0.729*
Students' education level					
Intercept	0.33	0.07	0.18	0.48	*<.001*
Tertiary	0.17	0.13	-0.10	0.45	*0.207*
Mixed	-0.23	0.15	-1.61	1.16	*0.340*
School Type					
Intercept	0.30	0.11	0.05	0.54	*0.021*
Public school	0.04	0.04	-0.05	0.13	*0.385*
Students' previous test or test preparation experience					
Intercept	0.25	0.06	0.11	0.38	0.002
Yes	0.44	0.14	0.10	0.79	0.020
Parts of students have relevant experience	0.03	0.13	-0.26	0.32	0.817
Not mention	0.18	0.15	-0.13	0.50	0.237
Random allocation					
Intercept	0.30	0.08	0.13	0.47	*<.001*
Yes	0.19	0.11	-0.05	0.42	0.111
Participant design					
Intercept	0.34	0.06	0.21	0.47	*<.001*
Within	0.17	0.15	-0.18	0.52	0.302
Comparison					
Intercept	0.33	0.07	0.18	0.47	*<.001*
Different test preparation methods	0.06	0.14	-0.25	0.37	0.681
Pre-test	0.23	0.21	-0.31	0.76	0.319
Publication type					
Intercept	0.39	0.07	0.25	0.53	*<.001*
Dissertation	0.00	0.16	-0.35	0.35	0.999

Report/Conference/Book chapter	-0.09	0.25	-0.8	0.61	0.734
Publication year					
Intercept	-5.64	4.73	-15.37	4.09	0.240
Year	0.00	0.00	0.00	0.01	0.220

4.3.1 Test specific characteristics

Nature of the test

According to the sub-group analysis, test preparation had significant effect for both high-stakes (g = 0.42, se = 0.08, k = 71, p < .001) and low-stakes tests (g = 0.27, se = 0.08, k = 26, p < .01). According to the meta-regression, there was no significant difference between test preparation effect for high-stakes tests and for low-stakes tests, b = -0.11, 95% *CI*: -0.28 to 0.07, p = .221. Similarly, no statistically significant difference was identified between test preparation effect for non-curriculum-based tests (such as SAT and IELTS), g = 0.29, se= 0.09, k = 41, p < .01, and for curriculum-based achievement tests (such as ACT), g = 0.36, se = 0.08, k = 56, p < .001; b = -0.07, se = 0.11, 95% *CI*: -0.28 to 0.14, p = .502.

Subject Domain

The average test preparation effects per subject domain have been listed in Table 7. The results revealed small differences between subject domains. The meta-regression results showed that the preparation for verbal tests (g = 0.28, se = 0.09, k = 29, p < .05) had significantly smaller effects compared with EFL tests (g = 0.52, se = 0.10, k = 37, p < .001), b = -0.25, se = 0.11, 95% *CI*: -0.48 to -0.03, p < .05. There was no significant difference between other domains and EFL (see Table 8).

Format of the test

The use of multiple-choice questions (MCQ) in the test did not have significant difference regarding the effect of test preparation, b = 0.11, se = 0.08, 95%*CI*: -0.08 to 0.30, p

= .202, indicating that the MCQ format is not a significant moderator of test preparation effect. Similarly, the use of short-response items in tests is not a significant moderator, b = -0.16, se = 0.08, 95%*CI*: -0.32 to 0.01, p = .060.

Test administration

The meta-regression indicated that the difference between test administration ways was not significant (b = 0.10, 95%*CI*: -0.05 to 0.26, p = .185), and the sub-group analysis showed that effects of test preparation intervention were similar regardless of whether the criterion tests were administered in real testing conditions (g = 0.40, se = 0.07, k = 43, p < .01) or administered as mocks (g = 0.29, se = 0.04, k = 54, p < .001).

4.3.2 Test preparation intervention design

Type of test preparation intervention

The sub-group analysis showed that commercial courses (k = 19) did not have a significant effect on test performance (g = 0.43, se = 0.19, p = .085). There was no significant difference between the effect of commercial courses and school courses (g = 0.36, se = 0.06, k = 19, p = .085) based on the meta-regression results, b = -0.11, 95% *CI*: -0.31 to 0.08, p = .246. Only two included studies examined the effect of students' self-preparation, g = 0.32, se = 0.02, p < .05. The meta-regression model showed no significant difference between the effect of self-study and the effect of commercial courses, b = -0.22, 95% *CI*: -2.61 to 2.17, p = .644. However, the result is not conclusive because the number of studies reporting student's self-study was not sufficient.

Breadth of test preparation intervention

Test preparation effects were moderated by the instructional breadth of test preparation intervention. When interventions targeted the same subject or area of the test, but their focus was not narrowed to exact tested content, the effect size was the largest, g = 0.38, se = 0.08, k

= 48, $p < .001$. Similarly, interventions focused solely on the test-specific content and format or drill on the exact same tested items also had a significant effect, $g = 0.36$, $se = 0.08$, $k = 37$, $p < .001$. In comparison, broad intervention focused on general knowledge or skill area had the smallest effect ($g = 0.13$, $se = 0.08$, $k = 12$, $p = .166$). The meta-regression showed that when compared to broad interventions, specific interventions and narrow interventions had significantly larger effects, $b = 0.19$, 95%*CI*: 0.04 to 0.34, $p < .05$; and $b = 0.14$, 95%*CI*: 0.01 to 0.29, $p < .05$, respectively.

Strategy use

The results of the sub-group analysis of the effect of test preparation strategies used during the interventions on student test performance can be found in Table 7. The meta-regression results suggested that test preparation had a significantly larger effect when interventions included test preparation management strategies ($g = 0.46$, $se = 0.08$, $k = 45$, $p < .001$) than those that did not ($g = 0.27$, $se = 0.07$, $k = 24$, $p < .01$), $b = 0.22$, 95%CI: 0.07-0.37, $p < .01$. Similarly, interventions in which students were taught test-taking skills ($g = 0.42$, $se = 0.07$, $k = 73$, $p < .001$) also had a significantly greater effect on students' test performance ($g = 0.43$, $se = 0.07$, $k = 75$, $p < .001$) compared with those that did not ($g = 0.16$, $se = 0.06$, $k = 22$, $p < .05$), $b = 0.21$, 95%CI: 0.02-0.39, $p < .05$. The use of other strategies, such as memorisation ($g = 0.34$, $se = 0.08$, $k = 73$, $p < .001$), socio-affective strategies ($g = 0.53$, $se = 0.14$, $k = 16$, $p < .01$), and broad knowledge development ($g = 0.37$, $se = 0.08$, $k = 56$, $p < .001$), also had significant impact on students' test performance according to the sub-group analysis, but when compared to interventions that did not report relevant strategy use, no significant differences were found (see Table 8 for meta-regression results).

Material use

The moderator analysis also examined that whether different material usage in test preparation intervention resulted in the variation of effect sizes. Interventions where students used commercial workbooks had a larger effect on the improvement of test performance (g = 0.47, se = 0.10, k = 49, p < .01) than those that did not (g = 0.23, se = 0.05, k = 48, p < .001), and the difference was significant according to the meta-regression (b = 0.28, 95% CI: 0.08-0.43, p < .05).

When test preparation interventions provided students with timed practice test which simulated the real test condition, the effect (g = 0.31, se = 0.09, k = 23, p < .05) was not significantly different to those that did not (g = 0.40, se = 0.07, k = 74, p <.001), according to the meta-regression results: b = -0.10, 95% *CI*: -0.35-0.15, p =.402. In addition, there was no difference between the interventions that involved sample items to practice (g = 0.38, se = 0.06, k = 75, p <.001) and those that did not (g = 0.30, se = 0.16, k = 22, p = .114), b = 0.00, 95% *CI*: -0.49-0.49, p =.998.

Besides, two included studies reported three effect sizes about the effect of test preparation interventions delivered virtually by computer-based programme. The estimate effect was g = 0.41, se = 0.04, k = 3, p = .061. The meta-regression showed that there was no statistically significant difference between the effect of test preparation interventions delivered in person (g = 0.37, se = 0.06, k = 94, p <.001) and those delivered virtually, b = 0.05, 95% *CI*: -0.71 to 0.82, p = .569.

Time and duration of the intervention

The meta-regression results showed that the number of contact hours was not a significant moderator of the effect, b = 0.00, 95% *CI*: -0.01 to 0.01, p = .644, nor was the duration of the intervention, b = 0.00, 95% *CI*: -0.02 to 0.02, p = .729.

4.3.3 Student-related characteristics

Education level

The test preparation effect was slightly larger for tertiary students (g = 0.50, se = 0.11, k = 31, p <.01) than for secondary students (g = 0.33, se = 0.07, k = 61, p <.001), but the difference was not significant, b = 0.17, 95% *CI*: -0.10 to 0.45, p = .207. Among studies that provided test preparation intervention for mixed-age groups, the test preparation effect was small and not significant (g = 0.10, se = 0.13, k = 5, p =.569).

Previous test experience

Among all included studies, 70% of cases reported whether participants had previous testing experience or test preparation experience on the specific test. The sub-group analysis showed that students with previous exposure to the test or test preparation seemed to benefit more from test preparation (g = 0.69, se = 0.13, k =12, p < .05); whereas for students without any previous testing experience, the test preparation effect was smaller, g = 0.25, se = 0.06, k = 37, p < .01). The difference between the two groups of students was significant according to the meta-regression, b = 0.44, 95%*CI*: 0.10-0.79, p < .05.

4.3.4 Study design characteristics

About 36% of included interventions used a randomized design. However, no statistically significant difference of the estimated effect sizes for studies with a random allocation and those without was identified in the meta-regression (b= 0.19, 95% *CI*: -0.05 to 0.42, p = .111).

A small proportion of included interventions (21 out of 97) utilized a within group design. The meta-regression revealed that the participant design method was not a significant predictor of the magnitude of test preparation effect, b= 0.17, 95% *CI*: -0.18 to 0.52, p = .302.

When the experimental and control conditions took different forms of test preparation, for example, intensive language test preparation courses vs. communicative language course (Nishitani, 2006; Pan, 2016), a significant effect was identified: $g = 0.34$, $se = 0.13$, $k = 20$, $p < .05$. This effect was similar with the interventions which compared the test preparation group (the experimental group) with no preparation group (the control group), $g = 0.33$, $se = 0.07$, $k = 59$, $p < .001$. The subsequent meta-regressions showed that the comparison condition was not a significant moderator of the effect, $b = 0.06$, 95% *CI*: -0.25 to 0.37, $p = .681$.

4.4 Publication bias

According to the analysis, the publication type, publication year, and the sample size were not significant moderators, indicating that there was less risk of publication bias. There were no significant differences between studies in dissertations or research reports and those that were published in journals (see Table 8 for the meta-regression result). Besides, the risk of publication bias was also evaluated by inspecting the funnel plot (see Figure 10), which gives an impression of the relationship between observed effects and standard error for asymmetry. For this meta-analysis, no systematic relationship was indicated by the funnel plot. The Egger's regression test (z=-0.67, $p = 0.5$) also suggested the risk of publication bias was judged to be low.

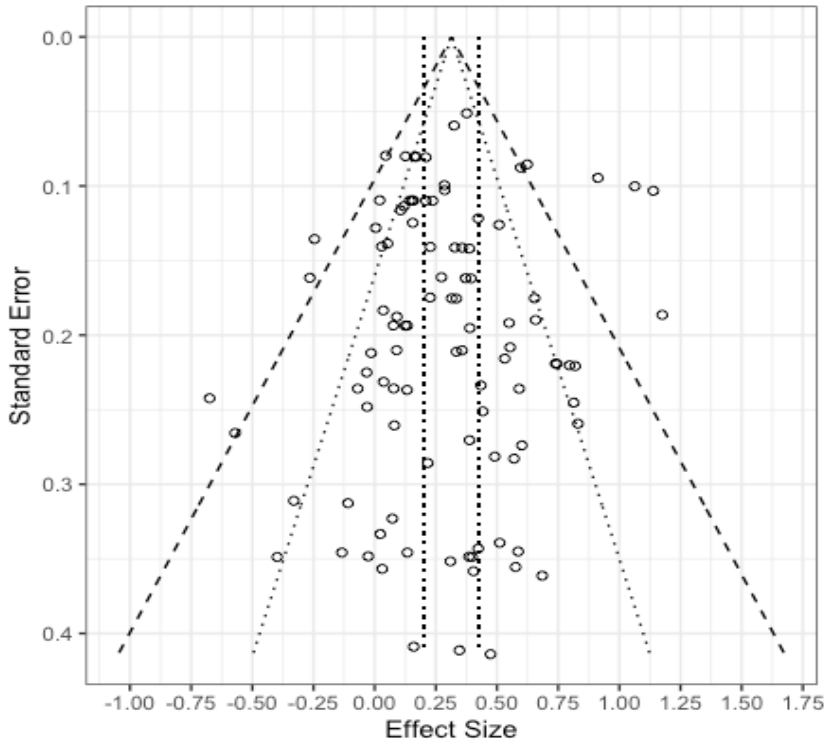


Figure 10 A funnel plot showing the relationship between standard error and observed effect size for the meta-analysis.

5. Discussion

The results of this meta-analysis show that test preparation has a significant small overall effect (g = 0.38) on test performance across a broad range of different domains and different tests in authentic educational contexts. Although the effect might be regarded as small to medium, according to Hattie (2012), this effect is very close to the hinge point (g = 0.4) of an average effect of educational interventions, which is typically considered practically significant. The findings also suggest that the test preparation effect is robust across different test conditions, for both large-scale high-stakes and low-stakes tests, curriculum-based and non-curriculum-based tests. EFL tests appear to be most susceptible to test preparation compared with other domains. We believe this might be explained by the higher predictability of EFL tests, as teachers and students have easy access to test-related materials

(such as workbooks and previous test papers), and they are familiar with the test format and content due to the high frequency of administration of some EFL tests.

The test preparation effect can be explained by four groups of explanations, as elaborated by Arendasy et al. (2016) and Lievens et al. (2007). First, test preparation may reduce construct-irrelevant factors, such as unfamiliarity and test anxiety, leading to an increase in test scores. Second, the development of test-wiseness and test-taking skills during test preparation increases the likelihood to solve test items correctly. Third, domain-specific knowledge or test-specific cognitive skills could be developed by test preparation. Fourth, generalisable latent cognitive ability and/or broad knowledge could be enhanced by test preparation intervention.

The moderator analysis of the observed test preparation effects provides some support for the first three explanations, as is discussed in the following. However, the fourth model was not supported by our meta-analysis. Our finding suggests that the test preparation effect is unlikely to transfer to other tests, which is in line with previous studies (see Arendasy et al, 2016; Coly, 2006; Freund & Holling, 2011; Koeing, Frey & Detterman, 2008) that found that preparing for a test by either coaching or repeated practice on sample items failed to improve general cognitive ability.

The first and second model of test preparation effects were corroborated by this meta-analysis. The analysis shows that students who had some previous test or test preparation experience possibly benefited from test preparation more than those who have not. This indicates how increasing test familiarity can help students get prepared for tests effectively. We also found that test preparation interventions that included instructions or practices on test-taking skills had a significantly greater effect than those that did not include test-taking skills training. However, test-taking skills such as guessing, wrong answer elimination, and time management may not be able to influence the underlying construct the test examined.

Such score increases might not represent improvements in the underlying construct highlighted by the test, thus undermining the validity of the interpretation of test scores.

In this meta-analysis, the breadth of test preparation intervention was categorised as narrow, specific, and broad, differentiating the alignment of teaching/learning content and the test. Findings indicate that interventions with a specific test-related focus and ones that even solely drilled students on tested content had a significantly larger effect on test performance than those that developed generalisable skills. Although test-oriented teaching and learning has received a lot of criticism, our findings confirm its effectiveness in increasing test scores and probably provides some empirical support for the first and third model suggested by Arendasy et al. (2016), according to which the increases of test scores could be attributed to test familiarisation and an increase of domain-specific knowledge or specific cognitive skills (Xie, 2013).

In fact, the framework of models of test preparation effect was mainly built upon evidence of retest effect on cognitive ability tests (Lievens et al., 2007). To some extent, the retest effect and test preparation effect shared similarities in terms of mechanisms. For example, both retesting and test preparation allow students to familiarise themselves with the test and test-taking skills. Additionally, retesting is an important and widely utilised way of test preparation. Therefore, it is plausible to explain the effects of test preparation using the framework of causes of retest effects. However, applying this framework directly to test preparation contexts has unavoidable limits due to the natural differences between retesting and test preparation. Compared to retesting, test preparation is a more complicated learning process which includes various cognitive and metacognitive components, where students interact dynamically with not only test papers and relevant materials, but also their teachers, peers, and even their parents.

This meta-analysis identifies that using workbooks is a significant moderator of the test preparation effect, as is developing test preparation management strategies during test preparation intervention, both of them significantly benefited test performance. These findings cannot be explained by the current framework drawn from retest effects (Arendasy et al, 2016; Lievens et al., 2007). Instead, learning and cognition theories provide hints as to how to explain the findings.

A test workbook typically contains an Introduction to the test structure, test format, sample test questions and answers, techniques, as well as some practice tests and review of knowledge content (see *Princeton Review SAT Prep* for example*)*, which provides scaffolding to help students build the schematic structure of the test and test content, which can further guide test preparation. A large number of studies confirmed the positive effect of scaffolding on educational achievements (Jitendra et al., 2009; Lin et al., 2911; Pawley et al., 2015). Nevertheless, we should be aware that scaffolding usually refers to 'a general teaching strategy and concept where educators provide guidance or other forms of support such as prompts, cues, instructions, information organization, or reference materials' (Perry et al., 2020, p.104), and previous studies primarily investigated the effect of scaffolding by manipulating classroom instructions, with no existing studies focused on the effect of scaffolding by using workbooks. More research is needed to support or reject this interpretation of the test preparation effect moderated by workbooks.

Our findings also reveal the value of incorporating metacognitive strategies into test preparation process, which we believe can increase test preparation effects. According to our findings, instructing students how to manage test preparation process through planning, goal setting, or analysing score rubrics and evaluating one's responses, etc., is a significant moderator. Despite inconsistent evidence in the literature on whether and how metacognitive learning strategies affect academic achievement (for reviews, see Boer et al., 2018; Dent &

Koena, 2016), our findings highlight the potential contributions of utilising metacognitive strategies on test preparation.

An unexpected but important finding of this meta-analysis is that the use of sample items and timed practice tests were not significant moderators of the effect of test preparation when analysed across studies. In previous meta-analyses of practice effect on cognitive ability tests (Hausknecht et al., 2007; Kulik et al., 1984; Scharfen et al., 2018), the use of sample items and practice tests have shown great impact on memory retention and the improvement of cognitive abilities. However, this meta-analysis finds little evidence of practice effect. The reasons might be that compared to cognitive ability tests, educational tests assess not only cognitive abilities, but also the understanding, evaluation and reflection of knowledge, as well as problem-solving ability and critical thinking ability. This might also be the reason why memorisation strategy is not a significant moderator in our study. Preparing for educational tests should be a comprehensive process that covers aspects of different skills other than knowledge retention. Besides, retest effects were only robust when using identical test forms in retesting sessions, while the retest effect for alternative forms of tests was found to be significantly smaller (Scharfen et al., 2018). In real educational contexts, it is rarely possible to practice test papers or items that are identical to those in real tests. We therefore infer that the practice effect can be smaller in educational tests. More quantitative and experimental studies are needed to test this inference.

Another unexpected finding of this study is that the effects of test preparation were significant for both high-stakes and low-stakes tests, but there was no difference between the two types of tests. This result seems to be unreasonable, as previous studies have shown that students were more motivated and tried harder in high-stakes tests (Knekta & Sundstrom, 2017; Liu et al., 2012). One of the possible explanations for this finding is that the results were derived from (quasi) experimental studies in which participants were given test

preparation intervention and required to prepare for the test regardless of the test's stakes. Such a design did not simulate real-world situations, where students might not study or pay less effort for low-stakes tests. Another reason could be the self-selection bias: students who opted to participate in the (quasi) experiments might be highly motivated and better engaged than those who did not participate. Thus, whether the test was high-stakes or low-stakes did not affect students' learning and thus did not moderate test preparation effects.

According to this meta-analysis, the effect of commercial test preparation courses did not significantly exaggerate additional test preparation courses provided by schools or universities. There has been a long debate on educational equity issues and commercial educational courses (shadow education). Some researchers were concerned that students from low SES groups might be disadvantaged because they cannot get equal access to commercial courses (Buchman et al., 2010; Zhang, 2011).

Lastly, compared to test preparation courses (coaching) provided by teachers, students' self-test preparation was under-investigated. Current experimental studies examined whether a specific instruction or teaching pattern was effective, regarding test preparation as teacher-driven activities, while students' potentials and initiatives were neglected. With more studies and better theory, it will be possible to identify features and dimensions of students' test preparation practices that are responsible for greater score gains.

6. Limitations

There were some limitations caused by the characteristics of some of the primary studies included in the meta-analysis. First, many studies provided insufficient information on the test preparation intervention, making it difficult to code some moderators. For example, most studies described the procedures of test preparation intervention very briefly without any detailed description of teachers' instructions. Therefore, we were unable to code the quality

of instructions and did not include this factor in the moderator analysis. Second, some included studies that evaluated the effect of a long-time test preparation intervention, i.e., a course that took 2 hours per week and lasted for 30 weeks, though they used a random allocation for participants. It can be problematic to attribute the score changes simply to the intervention without appropriate techniques for controlling other factors (such as other activities related to test preparation), thus the effect size might be inflated.

Besides, the findings in the meta-analysis are primarily based on studies carried out on college admission tests in the U.S. context. Although the findings indicate that the magnitude of the effects in the U.S. was similar in other contexts, and the difference between different tests was not significant, some caution should be taken when generalising results to other testing contexts, especially tests of other domains (i.e., science, medicine, foreign language except English) or in less investigated regions (i.e., regions in the Middle East, Southeast Asia, and Africa).

There were also limitations in terms of the interpretation of the results. There was no solid theory developed in the test preparation area to analyse its process and explain the effect, thus we interpreted the test preparation effect mainly based on the framework of retesting and some theories from cognitive science, which might not fit well and leave out of the scope other possible frameworks, such as self-regulation or instructional support. However, it is difficult to construct a valid framework for test preparation effect based on those studies without a rigorous design or when lacking precise descriptions on teaching and learning strategies used for test preparation. More rigorous and thoroughly described experiments on specific material use and strategy use during test preparation are needed.

7. Conclusion

This meta-analysis provides evidence of a significant positive effect of test preparation on test performance, particularly those test preparation interventions focused on relevant content and skills or drilled on tested content. It also highlights the moderating effects of the use of workbooks, managing own's test preparation process, teaching test-taking skills, and the development of socio-affective strategies. This study serves as a comprehensive quantitative synthesis of test preparation, contributing to the growing literature on the effectiveness of test preparation, and advancing our understanding of its strength and limitations. Further research on test preparation should include greater details of teachers' instructions and students' practices during the test preparation process, thus providing additional insights for using test preparation as effective learning and teaching activities.

References

* Alderman, D. L., & Powers, D. E. (1980). The Effects of Special Preparation on SAT-Verbal Scores. *American Educational Research Journal*, *17*(2), 239–251. https://doi.org/10.3102/00028312017002239

* Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement*, *35*(1), 31–47. https://doi.org/10.1111/j.1745-3984.1998.tb00526.x

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. American Psychologist, 36(10), 1086.

Appelrouth, J. I., Zabrucky, K. M., & Moore, D. W. (2017). Preparing students for college admissions tests. Assessment in Education: Principles, Policy & Practice, 24, 78–95. doi:10.1080/0969594X.2015.1075958

Arendasy, M. E., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests?. Intelligence, 55, 44-56.

* Bailey, D. R., & Judd, C. (2018). The effects of online collaborative writing and toeic

writing test-preparation on l2 writing performance. *Journal of Asia TEFL*, *15*(2), 383–397.

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication Bias: The Antagonist of Meta-Analytic Reviews and Effective Policymaking. *Educational Evaluation and Policy Analysis*, *34*(3). https://doi.org/10.3102/0162373712446144

* Barnett, J. E. (2000). Self-regulated reading and test preparation among college students. *Journal of College Reading and Learning*, *31*(1), 42–53.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research*, 60(3), 373– 417.

Blazar, D., & Pollard, C. (2017). Does Test Preparation Mean Low-Quality Instruction? *Educational Researcher*, 46(8), 420–433. https://doi.org/10.3102/0013189X17732753

* Bookman, A. B. (1981). *The effects of preparation for tests on standardized mathematics achievement test growth* (Doctoral dissertation, University of Connecticut).

* Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *Journal of Experimental Education*, *69*(2), 133–151. https://doi.org/10.1080/00220970109600653

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley and Sons. https://doi.org/10.1002/9780470743386

Bray, M. (2007). The Shadow Education System: Private Tutoring and its Implications for Planners, 2nd ed. UNESCO, International Institute for Educational Planning, Paris.

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS: 88. Chance, 14(1), 10–18.

Briggs, D. C. (2009). Preparation for College Admission Exams. 2009 NACAC Discussion Paper. National Association for College Admission Counseling.

* Brunner, M., Artelt, C., Krauss, S., & Baumert, J. (2007). Coaching for the PISA test. *Learning and Instruction*, *17*(2), 111–122. https://doi.org/10.1016/j.learninstruc.2007.01.002

Buchmann, C., Condron, C.D., & Roscigno, C.D. (2010). Shadow Education, American Style: Test Preparation, the SAT and College Enrollment. Social Forces, 89, 435–- 461.

* Bunting, B. P., & Mooney, E. (2001). The effects of practice and coaching on test results

for educational selection at eleven years of age. *Educational Psychology*, *21*(3), 243–253.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, *2*(2). https://doi.org/10.1177/2515245919847196

Domingue, B., & Briggs, D. (2009). Using Linear Regression and Propensity Score Matching to Estimate the Effect of Coaching on the SAT. *Multiple Linear Regression Viewpoints*, 35(1), 12-29.

* Ercanbrack, J. (1999). A Study of the Effect of Direct Test Preparation on the TOEIC Scores of Japanese University Students. *Tesl-Ej*, *3*(4).

* Farnsworth, T. (2013). Effects of Targeted Test Preparation on Scores of Two Tests of Oral English as a Second Language. *TESOL Quarterly*, *47*(1), 148–155. https://doi.org/10.1002/tesq.75

* Filizola, E. (2008). The effect of a test preparation course on the SAT scores of students at Saint Joseph academy. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, (May), 1264.

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test preparation in New Jersey: Inquiry- oriented and didactic responses. *Assessment in Education: Principles, Policy & Practice,* 11(1), 67–88.

* Gaye, Z. Z. (2001). The ACT Prep and its effects on students who participated in the Multicultural Educational Enrichment (MEE) program at Tennessee State University. Tennessee State University.

* Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy and Practice*, *14*(1), 75–97. https://doi.org/10.1080/09695940701272880

* Hardison, C. M., & Sackett, P. R. (2008). Use of writing samples on standardized tests: Susceptibility to rule-based coaching and the resulting effects on score improvement. *Applied Measurement in Education*, *21*(3), 227–252.

Hattie, J. (2008). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. New York, NY: Routledge.

Hausknecht, John & Halpert, Jane & Paolo, Nicole & Gerrard, Meghan. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. Journal of Applied Psychology. 92. 373-385. 10.1037/0021-9010.92.2.373.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta‐analysis. *Statistics in medicine*, 21(11), 1539-1558.

* Holmes, C. T., & Keffer, R. L. (1995). A Computerized Method to Teach Latin and Greek Root Words: Effect on Verbal SAT Scores. *Journal of Educational Research*, *89*(1), 47–50. https://doi.org/10.1080/00220671.1995.9941192

* Holthaus, R. A. (2008). *The effects of a high school coaching program on SAT-mathematics test scores.* Walden University.

* Justus, L. C. (2010). *Impact of a school-based test preparation course on ACT scores with consideration of cultural associations.* Union University.

* Khodabakhshzadeh, H., & Zardkanloo, R. (2017). The effect of mock tests on Iranian EFL learners' test scores. *International Journal of Education and Literacy Studies*, 5(3), 47-51.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. Psychological Bulletin, 95(2), 179.

Kulik, J. A., Kulik, C. L. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. American Educational Research Journal, 21(2), 435-447.

* Lane, K. L., Kalberg, J. R., Mofield, E., Wehby, J. H., & Parks, R. J. (2009). Preparing students for college entrance exams: Findings of a secondary intervention conducted within a three-tiered model of support. *Remedial and Special Education*, *30*(1), 3–18. https://doi.org/10.1177/0741932507314022

* Lee, J.-Y. (2019). Pedagogical Effects of Teaching Test-Taking Strategies to EFL College Students. *Reading in a Foreign Language*, *31*(2), 226–248.

Madaus, G. F. (1988). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, 65(3), 29–46.

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied*, *21*(4), 370–382. https://doi.org/10.1037/xap0000063

McGaghie, W. C., Downing, S. M., & Kubilius, R. (2004). What is the impact of commercial

test preparation courses on medical examination performance?. Teaching and learning in medicine, 16(2), 202-211.

* McMann, P. K. (1994). The effects of teaching practice review items and test-taking strategies on the ACT mathematics scores of second-year algebra students (Doctoral dissertation, Wayne State University).

* Mitchell, I., Nistor, N., Baltes, B., & Brown, M. (2016). Effect of Vocabulary Test Preparation on Low-Income Black Middle School Students' Reading Scores. *Journal of Educational Research and Practice*, *6*(1), 105–118.

Montgomery, P., & Lilly, J. (2012). Systematic reviews of the effects of preparatory courses on university entrance examinations in high school-age students. International Journal of Social Welfare, 21(1), 3–12. https://doi.org/10.1111/j.

Moore, R., Sanchez, E., & San Pedro, M. O. (2018). Investigating Test Prep Impact on Score Gains Using Quasi-Experimental Propensity Score Matching. ACT Working Paper 2018-6. *ACT, Inc.*

* Nishitani, A. (2006). Teaching grammar for the TOEIC test: Is test preparation instruction effective?.

* Pan, Y. C. (2016). Traditional and non-traditional test preparation practices: Learner performance and perspectives. *Electronic Journal of Foreign Language Teaching*, *13*(2), 170–183.

Powers, D. E., & Rock, D. A. (1999). Effects of coaching on SAT I: Reasoning test scores. Journal of Educational Measurement, 36(2), 93–118.

* Rao, C., McPherson, K., Chand, R., & Khan, V. (2003). Assessing the impact of IELTS preparation programs on candidates' performance on the GT reading and writing test modules (IELTS research reports Vol. 5, pp. 237-262). Canberra: IELTS Australia.

Raymond, D. J. W., & A. (2021). Comparison of Traditional and Game Enhanced Test Preparation in Technical Education. Grand Canyon University.

* Roberts, S. O., & Oppenheim, D. B. (1966). The effect of special instruction upon test preparation of high school students in Tennessee. ETS Research Bulletin Series, 1966(2), i-26.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal*

of Experimental Psychology: Applied, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Scharfen, J., Jansen, K. & Holling, H. Retest effects in working memory capacity tests: A meta-analysis. Psychon Bull Rev 25, 2175–2199 (2018). https://doi.org/10.3758/s13423-018-1461-6

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). Meta-analysis with R: Springer.

Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. *Journal of Educational Psychology*, *75*(1), 104–115. https://doi.org/10.1037/0022-0663.75.1.104

* Takallou, F., Vahdany, F., Araghi, S. M., & Tabrizi, A. R. N. (2015). The effect of test taking strategy instruction on iranian high school students' performance on english section of the university entrance examination and their attitude towards using these strategies. *International Journal of Applied Linguistics and English Literature*, *4*(6), 119–129. https://doi.org/10.7575/aiac.ijalel.v.4n.6p.119

Tanner-Smith, E.E., Tipton, E. & Polanin, J.R. Handling Complex Meta-analytic Data Structures Using Robust Variance Estimates: a Tutorial in R. J Dev Life Course Criminology 2, 85–112 (2016). https://doi.org/10.1007/s40865-016-0026-5

* Trenkic, D., & Hu, R. (2021). Teaching to the test: The effects of coaching on English-proficiency scores for university entry. Journal of the European Second Language Association, 5(1), 1–15. DOI: http://doi.org/10.22599/jesla.74

*Winke, P., & Lim, H. (2017). The Effects of Test Preparation on Second-Language Listening Test Performance. Language Assessment Quarterly, 14(4), 380–397. https://doi.org/10.1080/15434303.2017.1399396

* Zuman, J. P. (1988). The Effectiveness of Special Preparation for the SAT: An Evaluation of a Commercial Coaching School.

Author Contributions

Zhanxin Hao: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing-Original Draft.

Jo-Anne Baird: Conceptualization, Writing-Reviewing & Editing, Supervision

Yasmine El Masri: Conceptualization, Writing-Reviewing & Editing, Supervision

Kit Double: Methodology, Software, Writing-Reviewing & Editing

Chapter 5. Paper 2

This chapter displays the second research paper (Paper 2), which investigated the effects of using testing and restudy during test preparation sessions on students' performance of educational psychology tests. Whether effects of the two strategies varied across different conditions was also probed in this paper. This paper reported findings from three (quasi-) experiments. The results demonstrate that both testing (with feedback) and restudy were effective test preparation strategies and could improve students' performance on criterion tests administered immediately after test preparation sessions. However, only testing with feedback resulted in greater performance in delayed tests compared to the control group, while restudy did not. These significant effects were only observed in high-stakes conditions, where students prepared for high-stakes course tests. There was no significant test preparation effect in the low-stakes condition. Additionally, providing psychoeducation on metacognition and incorporating metacognitive monitoring and control tasks in test preparation practice did not improve the test performance significantly compared to the sole use of testing and restudy.

Effects of Using Testing and Restudy as Test Preparation Strategies on Educational Tests

Zhanxin Hao[1]; Jo-Anne Baird[1]; Yasmine El Masri[2]

[1] Department of Education, University of Oxford, Oxford, UK

[2] Ofqual, Coventry, UK

Corresponding details: zhanxin.hao@education.ox.ac.uk

Funding: This work was supported by Chinese Scholarship Council [grant number: 201808060181].

Effects of Using Testing and Restudy as Test Preparation Strategies on Educational Tests

Abstract

Students often prepare for tests by either restudying or taking practice tests that cover the content that will be assessed. We investigated the transfer effects of testing (both with and without feedback) and restudy strategies on test performance in an authentic test preparation context, as well as whether using metacognitive strategies improved the test preparation effect. We conducted three experiments in which students from a teacher training college in China reviewed educational psychology concepts using one of the strategies and took a post-intervention test immediately. In two of the three experiments, students sat a course test several days after the intervention. The results suggested that both testing with feedback and restudy were effective test preparation strategies with comparably similar effect sizes on test performance. Testing with feedback had a delayed effect on course test performance compared to the control group while restudy did not. These significant effects were only observed in high-stakes conditions, where students prepared for high-stakes course tests. There was no significant test preparation effect in the low-stakes condition. As indicated by greater post-test and final course test scores, the restudy effect was amplified by metacognitive prompts during test preparation.

Keywords: test preparation; testing effect; restudy effect; metacognitive strategies; test performance

Effects of Using Testing and Restudy as Test Preparation Strategies on Educational Tests

Introduction

Educational tests are widely used in educational systems around the world to evaluate students' progress and ability as well as provide important information for selection, admission, accountability and even policymaking. In many cases, preparing for tests has become an integral part of teaching and learning (Au, 2007). The main purpose of test preparation activities is to maximise scores on a specific test, hence many of them are intensive and cramming in nature, conducted within a short period of time before the test. Test preparation activities mainly include coaching courses provided by schools or commercial companies and students' self-preparation. Previous studies have reported positive effects of providing coaching courses on students' test performance (e.g., Kulik et al., 1984; Briggs, 2001). In comparison, students' self-preparation for the test and its effects are under-investigated. As test-oriented activities, test preparation usually focuses on a narrow scope of learning, thus it might have immediate effects on test scores but does not have long-term impacts on learning (Powers, 1985).

Our interest in the current research concerns the effect of students' self-preparation for a final course test in authentic education contexts, which have high stakes associated with the outcomes. Previous research has tended to be conducted in laboratory settings and involved memory for discrete pieces of information. We use materials from real classrooms to investigate the effects of two commonly used test preparation strategies on both immediate test performance and long-term retention of curriculum-based test items. We are also interested in the mechanism of test preparation effects and the conditions under which test

preparation works. Thus we examine whether test preparation effects differ when the stakes vary, as well as whether and how students' monitoring and control of their test preparation process affect the test preparation effects.

Effects of Testing and Restudy in Educational Contexts

One of the most commonly used methods for preparing for tests is to restudy the materials (Karpicke et al., 2009). Students report feeling more competent after restudying learning materials than taking practice tests (Janes et al., 2018). However, the effect of restudy does not appear to be impressive when compared with testing approaches, according to previous studies (Rawson & Dunlosky, 2011; Roediger & Butler, 2011). A large number of laboratory-based experiments has provided robust findings on the benefits of testing, which has been referred to as the *testing effect* or *retrieval-enhanced learning* (e.g., Carrier & Pashler, 1992; McDaniel et al., 2007; Roediger & Karpicke, 2006; Rowland, 2014). In typical laboratory experiments, students studied a set of materials (i.e., word pairs, text passages, video lectures), engaged in either retrieval or restudy conditions, then took a memory test immediately or after a delay of several days, weeks or months (Rawson & Dunlosky, 2011; Roediger & Butler, 2011). Educational researchers have applied similar designs in classroom settings, and their studies suggest that testing practices can promote students' learning across diverse educational stages, ranging from preschool (e.g., Fritz et al., 2007), primary schools (e.g., Goossens et al., 2014), secondary schools (e.g., Carpenter et al., 2009; Dirkx et al., 2014) to universities (e.g., Greving & Richter, 2018; Mcdaniel et al., 2007; Thomas et al., 2018) and different subjects (for recent reviews, see Agarwal et al., 2021; Schwieren et al., 2017; Yang et al., 2021).

Theoretical Explanation for Testing Effects

Previous studies indicated that testing could enhance learning, both directly and indirectly (Karpicke, 2017). Some laboratory studies suggested that students' scores increased by

simply engaging in testing practice without receiving feedback or subsequent restudy (Adesope et al., 2017; Kornell et al., 2009). Cognitive load and transfer appropriate processing theories are usually used to explain the direct effect of testing practice. Cognitive load theory posits that students have limited cognitive recourses, and the way in which these resources are expended during learning determines the effectiveness of learning activities (Sweller, 1988). According to this theory, testing practice can make recall automatic, thereby reduce the use of extraneous source (factors that do not contribute to learning) and save more capacity for germane source (relevant aspects of the task that can contribute to the learning) (Carpenter et al., 2020; Paas et al., 2003; Sweller et al., 1998). Subsequently, students are able to use additional capacity to process more learning tasks. The transfer appropriate processing framework suggests that memory improves when cognitive processes engaged during retrieval match the cognitive processes that were engaged when the material was encoded (Morris, Bransford, & franks, 1977). By engaging in testing practices, students match their initial processing to the type of processing required at the test (Roediger & Karpicke, 2006).

However, findings from recent classroom-based studies have questioned the direct effect of testing, suggesting that feedback provided after testing is the key for observing benefits of testing, which is regarded as the indirect effect of testing on learning (Roediger & Karpicke, 2006; Lipko-Speed, et al., 2014). Researchers demonstrated that feedback could improve students' elaboration of knowledge and help error correction on subsequent tests (Bangert-Drowns, Kulik, Kulik & Morgan, 1991; Butler & Roediger, 2008). Besides, the literature suggests that the indirect effects of testing may arise from effective monitoring of learning (Butler & Winne, 1995; Lipko-Speed, et al., 2014). There is growing evidence that testing has positive impacts on metacognitive monitoring (e.g., Tullis et al., 2013; Händel et al., 2020). For example, Cogliano et al. (2019) found that students possessed better monitoring

accuracy, less overconfidence and better performance for practice-tested items than untested items in college educational psychology classes. Barenberg and Dutke (2018) reported similar findings in a text-learning context with secondary school students, and proposed two mechanisms of the potential testing effect on metacognitive monitoring. First, testing might serve as an implicit cue that enables students to reflect upon the accessibility of information encoded in long-term memory from prior retrieval practice. Second, it is assumed that testing itself can promote elaboration of the encoded information, which would lead to a more solid knowledge base and further improve monitoring accuracy. However, Barenberg and Dutke (2018) did not provide students with feedback after testing, thus they did not discuss the potential role of feedback in the process. In a later study, Händel et al. (2020) assigned students to three groups: the testing group, the metacognitive training group, and the control group. Students in the testing group sat three mock tests and received written feedback one week after each test, while students in the metacognitive training group engaged in metacognitive training plus the mock tests. The training included psychoeducation, practiced judgements and feedback on judgement accuracy one week after testing. They found that repeated testing plus feedback improved not only test performance but also monitoring accuracy, and metacognitive training significantly added to the effects of repeated testing. However, the limitation of this research was that no information was collected on how the feedback was elaborated by students, making it difficult to determine the mechanism by which metacognitive training resulted in improved performance. It was unclear whether testing as an implicit cue improved monitoring accuracy, or whether students received feedback on judgements and performance, thereby elaborated more delicately on knowledge that was answered incorrectly or was misjudged.

The current research aims to provide empirical evidence for the above theoretical basis of testing effect. In the first two experiments, we compare effects of testing only with testing

plus feedback, on both immediate and delayed test performance to investigate the degree to which the testing effect is direct or if the effect is instead brought about through feedback in real test preparation contexts. The two experiments were identical, except Experiment 1 was conducted two days before the final course test and Experiment 2 was undertaken one week after the test. Thus, the variable of stakes of the test preparation settings was also manipulated. By comparing the results across Experiments 1 and 2, it is possible to determine whether the high-stakes context affects the degree of testing effect. In the third experiment, we further embed metacognitive training and practice – monitoring and control tasks – to testing and restudy practice to examine whether metacognitive training and practices during the test preparation process can optimise the effect of testing and restudy and lead to better test performance.

The Transferability of Effects of Testing and Restudy

Existing studies on testing effects have mainly investigated memory retention of discrete pieces of information that can be straightforwardly recalled, such as recall of foreign language vocabulary (e.g., Wheeler, Ewers, & Buonanno, 2003), text passages (Roediger & Karpicke, 2006), or picture recognition (e.g., Wheeler & Roediger, 1992). Studies in educational settings used a wider variety of materials compared to laboratory studies, including concepts or definitions from the course curriculum (e.g., Greving & Richter, 2018; Rawson et al., 2013; Carpenter et al., 2016), text arguments (e.g., Agarwal, 2019), diagrams (e.g., Freda & Lipp, 2016) and mathematics or science worked samples (e.g., van Gog et al., 2011). Most classroom-based studies still focused on memory retention: students were tested with identical or rephrased test questions that directly tapped into the learning materials presented in the intervention. For example, Roediger et al. (2011) provided students in a sixth-grade social studies course with materials about cultures and history from around the world. Students took classroom quizzes which included a random selection of half of the

target facts from their textbook and received immediate feedback. The researchers used exams to evaluate students' retention, including both free recall and multiple-choice questions of target facts. After a time-lag, students took final exams, which consisted of test questions from previous quizzes. This research found that students were more likely to correctly answer questions on exams if the target facts had been practiced on classroom quizzes.

Conclusions from these studies cannot be applied directly in real test preparation contexts where situations are more complicated. Testing items in educational tests are not always transparent and predictable, particularly in high-stakes-testing contexts. Typically, students learn test-related materials and practice sample items during test preparation, while encountering different items in formal tests. Preparing the exact same questions as those in official high-stakes educational tests in advance is usually associated with cheating behaviours. Therefore, it is important for educational researchers to carefully consider the transferability of testing effect between test preparation and formal educational tests, so as to better guide teachers and students' use of testing practices.

An increasing number of studies have explored the transferability of testing and restudy in different contexts. Some research has demonstrated that the effect of testing practices can transfer to similar or even different materials (e.g., Thomas et al., 2018). In their meta-analysis of testing effects in classroom-settings, Yang and colleagues (2021) found significant testing effects on not only identical and rephrased test materials, but also untested materials (g = 0.321, $p<.001$). However, some research showed contrasting findings, indicating that the effect of testing is not always significant over restudy when using different test items across intervention and criterion tests (e.g., Carpenter et al., 2020). In their second experiment, McDaniel et al. (2015) asked students to learn key research methods concepts and tested students using a different stem in questions in the final test. They found that the

repeated testing group did not outperform the repeated study group. Similarly, Peterson and Wissman (2018) examined the transfer of the testing effect on similar but not identical analogical problem-solving tasks and-found that participants in the testing group did not perform significantly better in the problem-solving task than the restudy group.

Pan and Rickard (2018) proposed a three-factor framework for transfer of testing effects. The first factor is response congruency between the initial and final tests, that the more similar the answers on the initial and final tests, the more likely the transfer. The second factor is whether elaborated retrieval practice is employed during training, including broad encoding methods, combining practice testing with restudy, feedback containing explanations or feedback that is presented for extended review. The third factor is the level of initial test performance, that transfer increases with higher accuracy on the initial test. The better one performs on the initial test, the more likely transfer will occur. This framework is mainly based on laboratory research, and more empirical evidence from authentic educational contexts is needed to investigate the theoretical and mechanistic basis for the transferability of testing.

In summary, whilst relevant findings in the literature are highly suggestive, they are insufficient to draw conclusions for practice in real education scenarios. In particular, the testing effect has been less explored using real life assessment tools. Therefore, the current research explored the effect of reviewing a series of curriculum materials by testing and restudy on real educational tests which target the same domain as the test preparation but use different test items. In this way we hope to simulate a more realistic test preparation context, and so contribute to a better understanding of the transfer effects of testing and restudy practices to prepare for tests in authentic educational settings.

Experiment 1: The Effect of Testing and Restudying During Preparation for a High Stakes Test

Method

Participants

Participants were 199 third-year students (68% female) from a teacher training college in China. Participants enrolled in a mandatory educational psychology course taught by the same teacher and took part in the study in exchange for course credit. Data were excluded for 19 participants who failed to complete the intervention due to technical problems or did not follow task instructions, and for 8 students who did not participate in the post-test, leaving 172 participants in total. The mean age of included participants was 18.3 years (*SD* = .85).

Design

This experiment used a one-factor design, with test preparation conditions manipulated between groups. Participants were randomly assigned to four conditions: testing with feedback, testing only, restudy, and control. The three experimental groups represented different test preparation strategies, while the control group represented no specific preparation condition (Table 9). The use of a control group offered a comparison with score gains from simply re-taking a test. The intervention was self-paced, and students' time use during the intervention was recorded.

Table 9 Experimental and Control conditions in Experiment 1.

Condition	Intervention
Testing with Feedback	Participants answered multiple-choice questions and short response questions from previous test papers or teaching slides. After submitting each answer, a message indicated whether the

	multiple-choice item was answered correctly, or exemplar answers of short response questions were presented.
Testing only	Participants answered multiple-choice questions and short response questions from previous test papers or teaching slides. After submitting each answer, no feedback was provided.
Restudy	Participants read definitions and statements of key concepts without testing or feedback.
Control	Participants solved Sudokus.

Materials

The intervention focused upon 24 concepts from six chapters of the educational psychology course: 'Psychological development and education of primary school students', 'Learning and learning theories', 'Learning strategies', 'Knowledge and Skills', 'Morality', and 'Teachers' Professional Development'. For each concept, one statement was created by summarizing the key information about the concept based on the textbook or teaching slides; and one test question from previous test papers was selected for each of the concept.

For example, a definition of *zone of proximal development* read:

ZPD is the difference between what a child can do alone and what a child can do with the help of a teacher, including a child's emerging skills.

The corresponding test question and correct answer read:

Question: What is the zone of proximal development?

Answer: ZPD is the difference between what a child can do alone and what a child can do with the help of a teacher, including a child's emerging skills.

The intervention materials consisted of either 24 summarizing statements (for restudy) or 24 practice questions (for testing). Note that 18 of the 24 test questions were multiple choice questions, while 6 were short-response questions. The researcher and the course

teacher jointly reviewed each statement, testing question and answer to ensure they were accurate.

Measures

The experiment utilised two measures.

Pre-test and post-test.

Two parallel test papers (A and B) developed by the course teacher were used as the pre-test and the post-test, both of which were designed in a similar structure to the course test and were compiled using test questions from previous final course tests. Test questions in the pre-test and the post-test were different from those used in the intervention. Each test paper contained 20 multiple-choice questions (two points for each correct answer), four definition recall questions (six points for each correct answer), and four analysis questions (nine points for each correct answer). All of the concepts reviewed in the intervention were assessed in each of the test papers, though with different test items. Randomly, half of the students in the experiment were assigned Test A for the pre-test and the other half used Test B for the pre-test. The groups performed similarly on the pre-test, with the two test papers assigned randomly, t (171) = -0.13, p = .896, indicating the difficulty levels of the two test papers were likely equivalent. Students who completed Test A in the pre-test were allocated Test B in the post-test and *vice versa*.

The course teacher created scoring protocols. Students' response on each multiple-choice question received either full points or zero. In terms of each short-response question, an ideal response was drafted and key answer points that would qualify for partial credit were identified by the course teacher. Students' responses did not need to be verbatim but should capture the correct meaning to receive credits. Based on scoring protocols, two independent markers (the first author and a research assistant) marked all scripts. Interrater reliability was

very good for each of the MCQ questions (all ICC > .85), and raters had high agreements when marking written answers of short-response questions (72% agreement). Disagreements were resolved by a third rater, who was one of the course teachers.

Final course test

A final course test was administered two days after the experiment. The final course test was constructed by two course teachers and assessed a broader range of curriculum content, covering topics from the whole textbook. Only those test questions covered by the topics in the intervention were selected for analysis and reported as the final course test results in this study, including 35 multiple-choice questions, six definition recall questions and four analysis questions (83 points in total). Since the pre- and post-tests were based on a total of 100 points, scores on the course test were converted by calculating the percentage of each participant's raw score on selected tested items out of the total score of 83 points and converting it to a score out of 100.

No identical test items from the pre-test and the post-test were used in the final course test. Two course teachers marked all scripts for the final test based on a scoring protocol, and they resolved score differences by discussion.

Procedure

The intervention materials were presented with the PsyToolkit platform (Stoet, G. 2010, 2017). One day before conducting the experiment, information sheets and consent forms were distributed to all participants. Participants were told that the experiment would prepare them for the final course test and investigate the effectiveness of different test preparation strategies. They were encouraged to do their best in the pre- and post-test and the intervention so that they might achieve better results in the course test. After participants signed and submitted the consent form, a 10-minute training session about how to use the online training

platform was provided. Following this, participants were informed of procedures of the experiment.

The experiment was conducted two days prior to the final course test. Participants went to four computer rooms to participate with their experimental group. All participants firstly took the pre-test, which lasted for 40 minutes. They had a 10-minute break after the pre-test. The intervention was self-paced with a maximum time of 60 minutes. During the intervention, participants in the two testing conditions answered test questions for each concept. Participants in the testing with feedback group received feedback immediately after they submitted their answers. Corrective feedback like 'Your answer is correct' or 'Your answer is not correct' were provided for multiple choice questions, and the test question was presented on the screen with the correct answer. For short response answers, the test questions and the answer showed on the screen after students submitted their answer. No feedback was provided for the testing without feedback group. Participants in the restudy group read the text relating to each concept. Each statement or test item was presented on one separate page, and the concepts were provided in the same sequence for the three experimental groups. Participants were permitted to take notes during the intervention. The control group solved sudoku puzzles. The post-test was provided after all students had finished the intervention. The high-stakes final course test data was collected two days after the intervention.

As the experiment was conducted in a real-life, high-stakes context, after the experiment, all materials were distributed to participants including those in the control group. The course teacher reiterated that students could use them as additional materials to prepare for the final course test. Students are free to study materials in advance of an examination in authentic settings and these materials simply represented another source of information.

Results

Table 10 summarises the mean scores and *SD*s for the four groups on all measures of Experiment 1. An a-level of .05 was used for all tests.

Table 10 Descriptive statistics of pre-test, post-test and final course test in Experiment 1.

Group	N	Pre-test		Post-test		Course test	
		Mean	SD	Mean	SD	Mean	SD
Testing with Feedback	46	36.83	12.38	47.17	11.16	53.85	8.18
Testing only	40	37.55	9.77	43.80	12.01	51.58	7.24
Restudy	41	38.73	10.88	45.32	12.10	53.56	8.79
Control	45	38.29	10.07	38.87	9.55	49.53	8.63
Total	172	37.83	10.79	43.77	11.55	52.20	8.43

Pre-test scores

To test whether the random assignment resulted in comparable groups, we first compared the pre-test scores between groups. With test preparation condition (testing with feedback, testing only, restudy, or control) as the independent variable and pre-test score as the dependent variable, a one-way analysis of variance (ANOVA) did not reveal any differences between conditions on pre-test scores, F (3, 167) = .26, p = .853, η_p^2= .005. Additionally, pre-test scores were positively correlated with post-test scores and final test scores (r = .23, p < .01 and r = .46, p < .001 respectively). To reduce error variance, we used the pre-test scores as a covariate in the analysis of post-test results and course test results.

Time on task

The intervention was self-paced, and individual student's time use on the intervention was recorded and analysed. A one-way ANOVA was used to examine the mean differences in time use among the three experimental conditions. A significant difference between the three groups was found, F (2,124) = 5.54, p < .01, η_p^2 = .082. Post Hoc Bonferroni tests confirmed

that the restudying group spent significantly more time during the intervention (*Mean* = 38.15, *SD* = 12.97) compared to the testing only group (*Mean* = 31.40, *SD* = 7.48), $p < .01$. The differences between the restudying group and the testing with feedback group (*Mean* = 33.35, *SD* = 6.87) were not significant. In order to control the variance caused by different time use, this variable was controlled as the covariate when comparing test performance between the three experimental groups.

Post-test scores

Pairwise tests showed that each experimental group achieved significant score gains from pre-test to the post-test, but there was no significant gain in the control group, t (45) = .47, p = .641. To compare the group differences on the post-test performance across the four groups, a one-way ANCOVA model with pre-test performance as the covariate was used. The results showed that there were significant differences in post-test scores between groups when the pre-test score was controlled as the covariate, F (3,167) = 5.14, p <.01, η_p^2= .084. A Post Hoc Bonferroni test confirmed that the testing with feedback group (*Estimated Marginal Mean*= 47.44, *SE* = 1.60) significantly outperformed the control group (*Estimated Marginal Mean*= 38.75, *SE* = 1.62), $p < .001$, as did the restudy group (*Estimated Marginal Mean*= 45.08, *SE* = 1.69), $p < .05$. The testing only group (*Estimated Marginal Mean*= 43.87, *SE* = 1.71) did not achieve significantly better scores than the control group, p = .189. The differences between the three experimental groups did not reach significance, ps > .05.

To control the variability in post-test performance that could be attributed to differences in time used by experimental groups, a one-way ANCOVA using pre-test performance and time use as the two covariates was conducted. The results showed that there was no significant group difference in post-test performance across the three experimental groups, F (2,122) = 1.05, p = .352, η_p^2 = .017. Post-hoc tests did not reveal any significant group difference, ps > .05.

Course test scores

We analysed students' course test performance on topics covered by the test preparation intervention. A one-way ANCOVA with test preparation conditions as the independent variable, final course performance as the dependent variable, and pre-test scores as the covariate, showed that there was a significant difference in mean scores between the four test preparation interventions conditions when controlling for pre-test scores, F (3,167) = 3.88, p < .01, η_p^2 = .065. Post Hoc Bonferroni tests revealed that there was a significant difference between the testing with feedback group (*Estimated Marginal Mean* = 54.22, *SE* = 1.06) and the control group (*Estimated Marginal Mean* = 49.36, *SE* = 1.07) on final test scores, p < .01. The restudy group (*Estimated Marginal Mean* = 53.22, *SE* = 1.12) and the testing only group (*Estimated Marginal Mean* = 51.68, *SE* = 1.14) did not achieve significantly better test performance than the control group, ps > .05.

A one-way ANCOVA controlling for pre-test performance and time use during interventions as two covariates was also performed. The results indicated that there was no significant group difference in the course test performance across the three experimental groups, F (2,122) = 1.19, p = .306, η_p^2 = .019. Post-hoc tests did not reveal any significant group difference, ps > .05.

Summary of Experiment 1

Findings of this experiment add to the widely held view in the literature that testing has a positive impact on students' learning when corrective feedback is given. This effect is also observable following a time lag (2 days in this experiment), as evidenced by the result of Experiment 1. The testing with feedback intervention produced significantly better test performance on both immediate post-test and the final course test. However, without feedback, testing itself produced very limited effects on students' learning and retention of

authentic educational materials. This inference was also supported by the fact that students in the control group who only took the pre-test and post-test did not acquire substantial score gains from the pre-test to the post-test. This finding is consistent with findings from the growing body of classroom-based literature, which also suggest that feedback is crucial for testing effects (e.g., Lipko-Speed et al., 2014; Kang et al., 2007; Smith & Karpicke, 2013).

Additionally, restudy had a significant immediate effect on students' retention though it did not benefit long-term learning as much as testing with feedback did. At first glance, this finding departs from the well-established conclusion in the literature that testing enhances retention relative to repeated study (see Rowland, 2014 for a meta-analysis). However, our experiment was conducted in a real test preparation context using authentic educational materials. The test questions in the pre- and post-tests and final course tests were relevant but not identical, requiring a certain amount of transfer of knowledge. Besides, test questions in this experiment examined students' understanding and application of knowledge rather than solely memory retention. In this case, the transfer effect of testing was not superior to restudy.

Experiment 2: Testing and restudying effect in a low-stakes context

Experiment 1 was conducted with students preparing for a high-stakes final course test. We are interested in learning more about whether the test preparation effect emerges when the test is low-stakes. Therefore, we repeated Experiment 1, with a different group of students taking part after the final course test. Students who participated in Experiment 2 were given a low-stakes post-test. Participants in Experiments 1 and 2 were from the same student cohort. They were assigned to the two experiments at random, thus allowing for a comparison between their post-test performances using the stakes of the test as an independent variable.

Method

Participants

Participants were 121 third-year students (70% female) in a teacher training college in China. Participants enrolled in a mandatory educational psychology course taught by the same teacher and took part in the study in exchange for course credit. Data were excluded for three participants who failed to complete the intervention or did not follow task instructions, and for three participants who did not participate in the post-test, leaving 115 participants in total. The mean age of participants was 18.3 years (*SD* = .91).

Design & Materials

The design and materials were identical to Experiment 1. This experiment was conducted one week after the final course test. Participants were randomly assigned to the same four conditions: testing with feedback, testing only, restudy, and control (see Table 9).

Measures

As in Experiment 1, two parallel test papers were used as the pre-test and post-test. Randomly half students in each conditions used Test A for pre-test while the other half was randomly selected to use Test B for pre-test. The students who completed Test A in pre-test were allocated Test B in the post-test and vice versa. The experiment was conducted one week after the final course test, thus students' performances on the course final test was collected as an indicator of students' prior knowledge. No delayed tests were used in this experiment to examine long-term retention but students' scores from the course test, taken before the experiment, were available for analysis.

Procedure

The procedures of this study were identical to Experiment 1.

Results

Table 11 presents the descriptive statistics for the final course test, the pre-test and post-test results of each group.

Table 11 Descriptive statistics of course test, pre-test, and post-test scores in Experiment 2.

Group	N	Course test		Pre-test		Post-test	
		Mean	SD	Mean	SD	Mean	SD
Testing with Feedback	32	48.53	11.15	50.03	17.85	55.75	17.47
Testing only	28	50.39	6.50	48.86	12.26	52.79	14.60
Restudy	31	49.35	8.68	51.06	14.31	57.97	17.00
Control	24	48.04	8.57	46.46	18.22	49.29	17.41
Total	115	49.10	8.90	49.28	15.68	54.28	16.76

Pre-test scores

The one-way ANOVA results showed that there was no significant difference of pre-test scores between the four test preparation conditions, F (3, 111) = .42, p = .741, η_p^2 = .011. This confirmed that comparable groups were created by the random assignment. Pre-test scores were positively correlated with post-test scores, r = .76, p < .001.

Time on task

The one-way ANOVA results indicated that there were significant group differences in time use during the intervention between the three experimental groups, F (2,88) = 10.19, p < .001, η_p^2 = .188. Post Hoc Bonferroni tests further revealed that the restudying group spent significantly less time during the intervention (*Mean* = 27.47, *SD* = 13.45) compared to the testing only group (*Mean* = 38.72, *SD* = 10.44) and the testing with feedback group (*Mean* = 38.40, *SD* = 8.92), ps < .001. The variable of time use was controlled as a covariate when post-test performance between experimental groups was compared.

Post-test scores

Using the pre-test score as the covariate, a one-way ANCOVA was performed to compare the post-test performance of the four groups. No significant difference in post-test scores between groups was found, F (3,110) = 1.10, p = .351, η_p^2 = .029. Adding time use as another covariate, the one-way ANCOVA model comparing post-test performance between the three experimental groups did not show any significant group difference, F (2,86) = 0.92, p = .403, η_p^2 = .021. No significant group difference was reported by post-hoc tests, ps > .05.

Cross-experimental analyses

Experiment 2 was conducted nine days later than Experiment 1, with all participants taking the final course test in between. A two (experiment sequence: Experiment 1 vs. Experiment 2) by four (test preparation condition: testing with feedback vs. testing only vs. restudy vs. control) mixed factorial ANOVA on course tests scores showed that there was a main effect of the experiment sequence on course test scores, F (1, 279) = 8.70, p < .01, η_p^2 = .031. Students in Experiment 1 achieved a significantly higher mean final course test score (*Mean* = 52.13, *SE* = .65) than those in Experiment 2 (*Mean* = 49.08, *SE* = .80). There was no main effect of test preparation condition, F (3, 279) = 1.34, p = .263, η_p^2 = .014. The interaction between test preparation condition and experiment sequence was not significant, F (3, 279) = .98, p = .402, η_p^2 = .010.

Though participants from the educational psychology course were randomly assigned to the two experiments, their prior knowledge might differ across the two experiments, as students in Experiment 2 might be more familiar with the course materials at the beginning of the intervention due to their preparation for the course test. We then performed a two (experiment sequence: Experiment 1 vs. Experiment 2) by four (test preparation condition: testing with feedback vs. testing only vs. restudy vs. control) mixed factorial ANOVA to examine students' prior knowledge level using the pre-test score as the dependent variable.

The results indicated that there was a main effect of the experiment sequence on pre-test scores, F (1, 279) = 50.78, $p < .001$, $\eta_p^2 = .154$. Students in Experiment 2 had a significantly higher mean score in the pre-test (*Mean* = 49.10, *SE* = 1.23) than those in Experiment 1 (*Mean* = 37.85, *SE* = 1.00), $p < .001$. There was no main effect of test preparation condition, F (3, 279) = .44, $p = .726$, $\eta_p^2 = .005$. The interaction between test preparation condition and experiment sequence was also not significant, F (3, 279) = .47, $p = .703$, $\eta_p^2 = .005$.

To further explore the extent to which the stakes of tests result in the variance of test preparation effects, a two (low *vs*. high stakes) by four (test preparation condition: testing with feedback *vs.* testing only *vs.* restudy *vs.* control) mixed factorial ANOVA was conducted to examine the score gain difference between the pre-test and post-test across different conditions and the two experiments. The results showed that only the main effect of group was significant, F (3, 279) = 3.13, $p < .05$, $\eta_p^2 = .033$. Pairwise tests revealed that the testing with feedback practice enhanced score gains compared with the control condition, t (146) = 6.88, $p < .01$. There was no significant difference between other experimental groups and the control group. Although the high-stakes test preparation groups had a numerically larger score gain between the pre-test and the post-test compared with the low-stakes test preparation groups, the main effect of stakes of tests was not significant, F (1, 279) = .51, $p = .476$, $\eta_p^2 = .002$. The interaction between test preparation condition and the stakes of test was also not significant, F (3, 279) = .976, $p = .404$, $\eta_p^2 = .010$.

Summary of Experiment 2

The key finding of Experiment 2 was that in the low-stakes test preparation context, participating in test preparation did not have significant benefits on students' test performance. Students in the restudy group in this experiment spent the least time, on average, on the intervention and achieved the largest score gains after the intervention

compared with other groups, though the score gain differences were not significant. Besides, testing (either with or without feedback) did not yield significant test performance compared with the control group.

The cross-experimental analysis showed that students in Experiment 1 had, on average,better performance in the course test than those in Experiment 2, which further confirmed the effect of test preparation intervention before the high-stakes test. Besides, the testing with feedback group significantly outperformed the control group in the post-test cross the two experiments, indicating that the testing with feedback is an effective test preparation strategy.

Experiment 3 The Effect of Using Metacognitive Strategies with Restudy and Testing for Test Preparation

In Experiment 3, we partly replicated the design of Experiment 1 to provide verification of those findings, comparing the effect of using restudy and testing as test preparation strategies on an immediate test and a high-stakes course test one week after. In this experiment, we removed the testing only condition as it appeared to be the least beneficial strategy according to the findings in the previous two experiments. Besides, in authentic educational contexts, formative testing is often accompanied by the provision of feedback of some kind, while using testing without any feedback is not common.

In addition, we investigated the effect of integrating metacognitive training and metacognitive tasks during test preparation on the test performance to explore whether the testing and restudy effect could be enhanced, which might further develop our understanding of the potential mechanisms for the test preparation effects. The assumption was that in the metacognitive conditions, students would make more effort on tasks for which they are unconfident, thus leading to better regulation of their test preparation and yielding better test

performance. To investigate this assumption, we collected and analysed data from students' study behaviours in the experimental groups during the intervention, including learning attempts of each topic and the overall time use.

Method

Participants

Two hundred and sixty-eight students (72% female) from a teacher training college in China consented to participate in the study. The educational psychology course was one of the compulsory courses in the school and all participants took the course provided by the same teacher. Participants were all third-year students from five classes. Three classes were randomly chosen to receive the intervention, while two classes served as the control group. In each treatment class, students were randomly allocated to one of four conditions. Data were excluded for nine participants who failed to complete the intervention or did not follow task instructions, and for two participants who did not participate in the post-test, leaving 257 participants in total. The mean age was 17.34 (SD = .21).

Design

This study followed a quasi-experimental pre-post-design. A 2 × 2 factorial design was used, in which both test preparation practices (test preparation: restudy *vs*. testing with feedback) and metacognitive training and practices (metacognition: yes *vs*. no) were manipulated between subjects. As shown in Table 12, the design resulted in four intervention conditions in which participants were randomly assigned. In addition, students from two classes served as the business-as-usual control group. When the experimental groups received the intervention, students in the control group watched a 40-minute video lecture about educational psychology theories and teachers' professional development, which was part of the curriculum but not related to the test content.

Table 12 Experimental and Control conditions in Experiment 3.

Group name	Condition
Testing with feedback	Participants answered multiple-choice questions and short response questions from previous test papers or teaching slides. After submitting each answer, feedback was provided.
Testing with feedback plus metacognition	Participants received a 10-minute metacognition training session before the intervention. During the intervention, participants indicated their confidence in mastery of the topic by rating from 1 (very unconfident) to 10 (very confident). Then they answered multiple-choice or short response questions related to the topic. After answering the test question, they rated their confidence in mastery again. Corrective feedback was provided after submitting the answer and the judgements. Participants were asked to decide whether to re-answer this item or review the next topic
Restudy	Participants read definitions and statements of key concepts.
Restudy plus metacognition	Participants received a 10-minute metacognition training session before the intervention. During the intervention, participants firstly indicated their confidence in mastery of the topic by rating from 1 (very unconfident) to 10 (very confident). Then they read a statement of the topic. After reading, they rated their confidence again. After submitting the judgements, participants were asked to decide whether to re-read the topic or review the next topic.
Control	Participants watched a lecture on teacher professional development.

Materials

This intervention focused upon 16 concepts from the educational psychology course. This study was carried out at the end of the first term of the course, when students had learned four chapters from the whole curriculum. Therefore, four concepts were chosen from each of the 4

chapters: ‘Psychological development and education of primary school students’, ‘Learning and learning theories’, ‘Learning strategies’, and ‘Knowledge and skills’. For each concept, one statement and one test question were created by summarizing the key information about the concept based on the textbook or teachers teaching slides. The researcher and the course teacher discussed each statement and each test question to make sure they were stated properly and accurately. The intervention materials consisted of either 16 summarizing statements (for restudy groups) or 16 practice questions (for testing groups). Each statement or test item was presented on one separate page, and the presentation order of concepts was consistent.

Measures

Three types of measures were used in this experiment.

Pre-test and post-test

Two parallel test papers (Test paper C and Test paper D) developed by the course teacher were used as the pre-test and the post-test, both of which were designed in a similar structure to the course test. All of the concepts in the intervention were covered in each of the test papers, though with different test items. Each test paper contained 10 multiple-choice questions, 2 definition recall questions, and 4 critical analysis questions. Test papers used in this experiment were shorter than tests used in Experiments 1 and 2, due to the constraints of course time. Randomly, half of the students in the experiment were assigned Test C for the pre-test and the other half used Test D for the pre-test. Students who completed Test C in the pre-test were allocated Test D in the post-test and *vice versa*. The groups performed similarly on the pre-test, with the two test papers assigned randomly, $t(171) = -1.01$, $p = .314$, indicating the difficulty levels of the two test papers were likely equivalent.

Scoring and score resolution was conducted as in the previous experiments. Interrater reliability was very good for each of the multiple-choice questions (all ICC > .85). Markers had high agreements when marking written answers of short-response questions (86% agreement). Disagreements were solved by discussion.

Metacognitive judgements

During the pre- and post-tests, students were required to make item-specific judgements on each test item, indicating the extent to what they were confident that their answer was correct ranging from 1 (very unsure) to 10 (very sure). The absolute accuracy of students' judgement was calculated, using the formula shown below (Schraw, 2009):

$$\text{Absolute Accuracy Index} = \frac{1}{N}\sum_{i=1}^{N}(c_i - p_i)^2$$

where c_i corresponds to a metacognitive judgement, p_i corresponds to a performance score, and *N* is the number of judged items. To note, performance scores for multiple choice items were converted to either 1 (incorrect) or 10 (correct), while scores for short-response items were converted to a 10-point scale based on the percentage of scores earned out of the total score for each item.

Final course test

A high-stakes final course test was administered seven days after the experiment, which was longer than the pre- and post-test, composed of 40 multiple-choice questions, 7 definition recall questions, and 4 analysis questions. The course exam was based on these four chapters that were the focus of the intervention, but it consisted of a different set of test questions than the pre- and post-tests. Therefore, raw scores of the course test were used in the analysis. The course test was marked by two course teachers independently and marking differences were solved by discussion.

Procedure

One day before conducting the experiment, information sheets and consent forms were distributed to all participants. Participants were told that the experiment investigated the effectiveness of different test preparation strategies and would help them to prepare for the course test. After participants signed and submitted the consent form, they were informed of which group they were allocated to, as well as procedures of the experiment.

This study was conducted in classrooms during a special test preparation session of the course one week before the final course exam. Five classes participated in the study over two days. The researcher requested that test materials and training materials should not be disclosed to others until the end of the experiment. Students who participated in the intervention were encouraged to do their best in the online intervention as well as in the pre- and post-test so that they might achieve better results in the course test. Before the intervention, a 10-minute training session about how to use the online training platform was provided. Students practiced several example questions to familiarise themselves with the task during the training. Additionally, students in the two monitoring and control groups received another 10-minutes psychoeducation on the importance of metacognition as well as how to use metacognitive strategies to regulate their learning. They read several sentences about the importance of accurately monitoring learning and allocating study efforts to weak points in their learning.

After the training session, the pre-test was administered. Participants in each class were randomly allocated to one of the four conditions, and randomly half students in each condition used Test C for the pre-test, and the other random half students used Test D for the pre-test. The pre-test lasted for 25 minutes, and the intervention started immediately after the pre-test. The intervention was presented on the PsyToolkit website and students could access it via laptops, tablets or mobile phones. The practice was self-paced, and the researcher

indicated to students the remaining time every ten minutes. After the intervention, students took a 10-minute break before they took the post-test, which was conducted in the same manner as the pre-test. Students who used Test C in the pre-test were allocated Test D and vice versa.

Results

Study behaviours

In this experiment, study behaviours of the students in four experimental groups were recorded and analysed. The descriptive information of students' learning attempts and the average time use in each experimental group was illustrated in Table 13. We first performed 2 (test preparation strategy: restudy *vs.* testing) × 2 (metacognition: yes *vs.* no) mixed ANOVA models to investigate the group difference in study behaviours between groups.

Table 13 Study behaviours of students in experimental groups in Experiment 3.

Group	N	Study attempts		Time use	
		Mean	SD	Mean	SD
Testing with feedback	39	16.26	.677	31.07	4.97
Testing with feedback plus metacognition	38	20.53	3.28	33.44	4.79
Restudy	39	16.05	.223	26.21	7.24
Restudy plus metacognition	37	17.32	1.56	31.67	5.49

Time on task

The two-way ANOVA on time use revealed that there was a significant main effect of the metacognitive strategy on time use, F (1,149) = 17.88, $p < .001$, $\eta_p^2 = .107$. Students in metacognitive groups (*Mean* = 31.68, *SD* = .69) spent more time on average on self-preparation than others (*Mean* = 29.57, *SD* = .69), $p < .001$. The main effect of test

preparation strategy was also significant, F (1,149) = 12.85, p < .001, η_p^2 = .079. The pairwise test showed that students in testing conditions (*Mean* = 32.26, *SD* = .65) spent a longer time on the task than those in restudy conditions (*Mean* = 28.94, *SD* = .66), p < .001. The interaction between the two factors was not significant, F (1,149) = 2.80, p = .096, η_p^2 = .018. However, post-hoc tests showed that testing with feedback group spent significantly more time on the task than those in the restudy group (*Mean difference* = 4.86, *SE* = 1.29, p < .001), and the restudy plus metacognition group also spent significantly more time on the task than the restudy group (*Mean difference* = 5.45, *SE* = 1.31, p < .001).

Study attempts

In terms of students' study attempts, the mixed ANOVA results showed that there was a significant main effect of metacognitive training and tasks, F (1, 149) = 86.66, p <.001, η_p^2 = .368. Students in the two metacognitive groups showed significantly more study attempts (*Mean* = 18.93, *SD* = .213) than those without metacognitive interventions (*Mean* = 16.15, *SD* = .208), p < .001. There was also a significant main effect of test preparation strategy, F (1, 149) = 32.74, p <.001, η_p^2 = .180. The pairwise test showed that compared to the two restudy groups (*Mean* = 16.69, *SD* = .211), students in the testing with feedback groups had significantly more study attempts (*Mean* = 18.39, *SE* = .210), p <.001. The interaction between the two factors was also significant, F (1,149) = 25.33, p < .001, η_p^2 = .145. Post-hoc tests showed that testing with feedback plus metacognition group had significantly more study attempts than the testing with feedback group (*Mean difference* = 4.27, *SE* = .42, p < .001) and the restudy plus metacognition group (*Mean difference* = 3.20, *SE* = .43, p < .001), while the restudy plus metacognition group had significantly more study attempts than the restudy group (*Mean difference* = 1.27, *SE* = .42, p < .01).

Confidence ratings

During the intervention, students in the two metacognition groups rated their confidence before and after learning each topic. A series of independent sample t-tests were used to compare the group difference in students' confidence levels. The results showed that the restudy plus metacognition group reported higher confidence level, on average, before learning each topic (*Mean* = 6.60, *SD* = 1.43) than the testing plus metacognition group (*Mean* = 5.82, *SD* = 1.46), $t = 2.33$, $p < .05$. After learning the topics, the two groups reported similar levels of confidence, *Mean* = 6.66 (*SD* = 1.47) and *Mean* = 6.49 (*SD* = 1.48) separately, $t = .50$, $p = .62$. In terms of the changes in confidence, the testing plus metacognition group had significantly greater gain in confidence after learning (*Mean* = .67, *SD* = .66) than the restudy plus metacognitive strategies group (*Mean* = .06, *SD* = .19), $t = 5.38$, $p < .001$.

Relationships between group, confidence ratings and study attempts

We further explored the relationship between students' confidence rating on each topic and study attempts by performing grouped linear regression models. Using students' study attempts for each topic as the dependent variable and the confidence rating of each corresponding topic before learning as the predictor, the results of linear regression models showed that in the testing plus metacognitive strategies group, initial confidence could significantly predict the study attempts, $b = -0.02$, $SE = .005$, $p < .001$, indicating that the less confident students were, the more likely they studied the topic repeatedly. However, in the restudy plus metacognitive strategies group, the initial confidence rating was not a significant predictor for the study attempts, $b = -0.0004$, $SE = .005$, $p = .935$.

Considering that students were randomly allocated to the two experimental groups, and they learned 16 topics during the intervention, there might be variation between groups and between the topics. Therefore, we used a mixed effect model where the initial confidence and group are fixed effects, as we want to make a statement about these specific levels. We

treated students and topics as random, as we want the fixed effects to be population average effects and we were interested in the variation between different students and different topics. We used the model:

$$Study\ attempts = Intercepts + Initial\ Confidence + Group + Random\ Effects\ of\ Students\ (nested\ in\ group) + Random\ Effects\ of\ topics$$

Note that in this model, restudy plus metacognition group was specified as 0, while testing plus metacognition group was specified as 1. Data were analysed in R using lme4 and lmerTest packages.

As shown in Table 14, there was a significant main effect for group (b = 0.29, se = .07, p < .001), indicating that the testing plus metacognition group associated with more study attempts. The main effect for confidence rating was not significant, b = -0.005, se = .01, p =.504. The interaction between group and confidence rating was significant, b= -0.02, se = .01, p < .05. Figure 11 plots the interaction between the two variables, demonstrating that in general when students were less confident, they tended to use more study attempts. The impact of confidence level on number of study attempts seemed to be greater in the testing with metacognition group, rather than the restudy plus metacognition group.

Table 14 Mixed effect model results of the learning attempt.

Fixed effects	***b***	***SE***	***t-value***	***p***
Intercept	1.11	0.06	19.09	<.001 ***
Initial Confidence	-0.00	0.01	-0.67	.504
Group	0.29	0.07	4.26	<.001 ***
Confidence:Group	-0.02	0.01	-2.17	.030 *
Random effects	***Variance***	***SD***		
Student:Group	0.01	0.12		
Topic	0.06	0.11		
Residual	0.12	0.35		

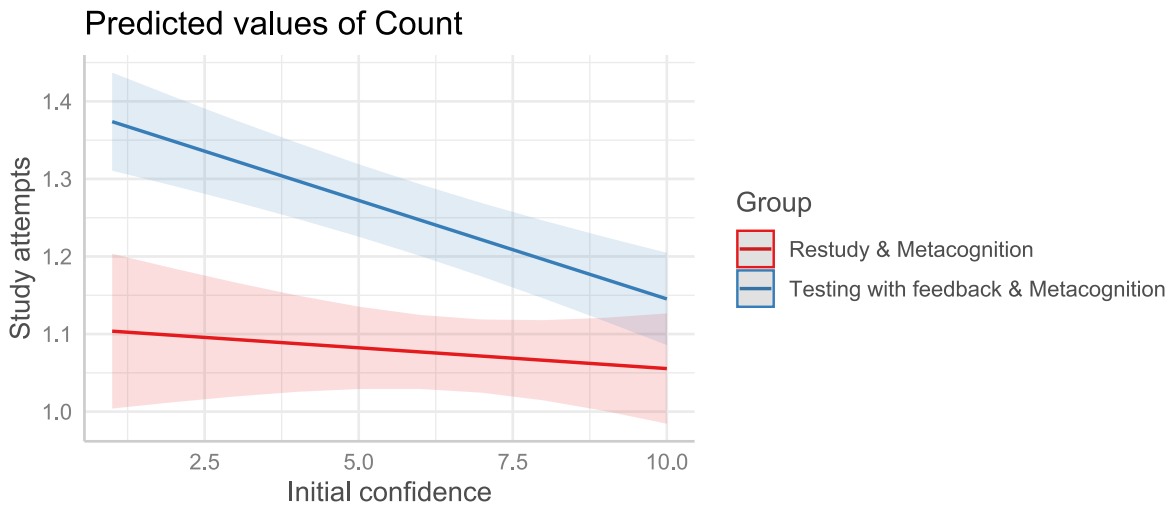


Figure 11 The interaction between group and confidence on study attempts.

Students' test performance

Table 15 presents the descriptive statistics for the pre-test, post-test, and final course test results of each group, as well as the absolute accuracy of students' judgement for the pre- and post-tests.

Table 15. Descriptive statistics of pre-test, post-test, and course test scores in Experiment 3.

Group	N	Pre-test scores		Pre-test absolute accuracy of monitoring		Post-test scores		Post-test absolute accuracy of monitoring		Course test scores	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Testing with feedback	39	33.77	7.50	0.15	0.05	42.85	10.63	0.14	0.06	65.21	10.35
Testing with feedback plus metacognition	38	35.13	9.70	0.15	0.04	44.71	12.19	0.14	0.06	62.95	13.02
Restudy	39	33.49	10.11	0.13	0.05	42.00	11.50	0.14	0.05	60.21	15.21
Restudy plus metacognition	37	34.08	9.27	0.14	0.05	43.71	8.54	0.13	0.05	61.49	12.91
Control	104	36.21	9.52	0.13	0.05	36.28	10.19	0.15	0.05	57.40	14.02
Total	257	34.92	9.32	0.14	0.05	40.47	11.08	0.14	0.05	60.42	13.62

Pre-test scores

The one-way ANOVA did not find any difference between the five conditions on pre-test scores, $F(4, 253) = .96$, $p = .431$. Pre-test scores were positively correlated with post-test scores and final test scores ($r = .55$ and $.59$, $ps < .001$). To reduce error variance, we used the pre-test scores as the covariate in the analysis of post-test performance and final course test performance.

Post-test scores

A 2 (test preparation strategy: restudy *vs*. testing) × 2 (metacognition: yes *vs*. no) factorial design resulted in four experimental conditions. To understand how the mean of post-test performance changed according to the levels of two independent variables, a two by two mixed factorial ANCOVA was conducted using post-test score as the dependent variable, and pre-test score as the covariate. The time use in each experimental group was also controlled as another covariate to reduce the potential confounding effect, as it positively correlated with the post-test results ($r = .28$, $p < .001$). The results showed that there was no significant main effect of test preparation strategy, $F(1,147) = 0.22$, $p = .640$, $\eta_p^2 = .001$, with testing (*Estimated Marginal Mean* = 42.99, *SE* = .97) and restudy (*Estimated Marginal Mean* = 43.65, *SE* = .97) producing similar effects on post-test performance, $p = .640$. Similarly, there was no significant main effect of metacognitive strategies, $F(1,147) = .002$, $p = .963$, $\eta_p^2 = .000$. Students using metacognitive strategies plus test preparation performed similarly with normal test preparation conditions (Estimated Marginal Mean = 43.35 vs. 43.28), $p = .963$. No significant interaction was found between the two independent variables, $F(1,147) = .010$, $p = .921$, $\eta_p^2 = .000$.

To compare the post-test performance between the four experimental groups and the control group (the no test preparation condition), a one-way ANCOVA was further performed using conditions as the independent variable, post-test performance as the

dependent variable and pre-test performance as the covariate, showing that there was a significant difference between groups, $F(4,251) = 16.39$, $p < .001$, $\eta_p^2 = .207$. Post hoc Bonferroni tests showed that compared with the control condition (*Estimated Marginal Mean= 35.37*, *SE* = .82), each experimental group performed significantly better in the post-test, *Estimated Marginal Mean* = 43.65 (*SE* = 1.34), 44.56 (*SE* = 1.35), 43.00 (*SE* = 1.34), 44.55 (*SE* = 1.37) respectively for testing with feedback, testing with feedback plus metacognitive strategies, restudy, and restudy plus metacognitive strategies groups, *ps* < .001.

Final course test scores

We analysed students' performance on course tests among those topics covered by the intervention by the 2 (test preparation strategy) × 2 (metacognition) mixed factorial ANCOVA with the pre-test results and time use controlled as the covariate. The results were similar to the ANCOVA model on post-test performance. No significant main effect of test preparation strategy was identified, F (1, 147) = .89, p = .347, η_p^2 = .006, nor did the metacognitive strategy, F (1, 147) = 1.70, p = .195, η_p^2 = .011. There was no significant interaction between the two independent variables, F (1, 147) = 1.14, p = .287, η_p^2 = .008.

We further conducted a one-way ANCOVA to compare students' course test performance between the experimental groups and the control group, using pre-test score as the covariate. The results showed that there was a significant difference between the groups, F (4, 251) = 7.98, $p < .001$, $\eta_p^2 = .113$. The post-hoc tests showed that the testing with feedback group received the relatively highest course test scores (*Estimated Marginal Mean* = 66.24, *SE* = 1.68) and significantly outperformed the control group (*Estimated Marginal Mean* = 56.24, *SE* = 1.03), $p < .001$; followed by the testing plus metacognition group (*Estimated Marginal Mean* = 62.76, *SE* = 1.70, $p < .05$), and the restudy plus metacognition group (*Estimated Marginal Mean* = 62.51, *SE* = 1.73, $p < .05$). The restudy only group did

not have significantly higher scores (*Estimated Marginal Means* = 61.50, *SE* = 1.54) than the control group, $p = .084$.

Accuracy of metacognitive judgements

Data on the absolute accuracy of metacognitive judgements are presented in Table 15. To investigate whether different test preparation strategies, especially the incorporation of metacognition during the intervention could improve students' monitoring accuracy, an ANCOVA was conducted to explore the variance of absolute accuracy of metacognitive judgements in the post-test among groups, while the absolute accuracy of judgements in the pre-test was controlled as the covariate. The results indicated that no significant mean difference existing between groups, $F(4, 251) = 0.63$, $p = .640$, which implied that in this experiment, both restudy and testing practices, whether with or without monitoring and control tasks, did not improve the absolute accuracy of students' monitoring in the post-test.

Summary of Experiment 3

Findings from this experiment indicate a similar pattern of test preparation effect as in Experiment 1; that for the immediate test, both testing with feedback and restudy are effective test preparation strategies in high-stakes test preparation contexts compared to no test preparation. However, the restudy only did not show significant effect on the delayed course test compared to the control.

In this experiment, an additional variable—metacognition—was incorporated to examine whether students' monitoring and control of their test preparation could foster the effect of test preparation strategies. We found that using a testing strategy in conjunction with metacognitive training and tasks for test preparation elicited more testing attempts and increased students' confidence of learning than using restudy together with metacognitive training and tasks. However, the two metacognition groups failed to yield larger test

preparation effects compared to the two normal test preparation groups in our experiment when scores on the post-test or final course were compared. One possible reason might be that the training and practice of monitoring and control in this experiment did not improve students' monitoring accuracy and students were not able to regulate their test preparation more effectively, thus there was no additional improvement in test performance.

General discussion

In this research, our primary goals were to investigate the effectiveness of testing and restudy as test preparation strategies in both high-stakes and low-stakes testing contexts, and whether the test preparation effect could be optimised when students monitor and control their learning.

Testing with feedback and restudy are effective test preparation strategies

Results of the three experiments demonstrated that testing with feedback is an effective test preparation strategy for high-stakes tests, as it significantly benefited students' performance on an immediate test and delayed course test compared with the control condition. However, the testing only practice did not yield better test performance in both high-stakes and low-stakes contexts. Since we used different test items in the pre-test, the post-test, and the course test, the results suggested that feedback is the key for enabling the transfer of test-enhanced learning in the authentic educational context. This finding provides empirical support for the transfer framework proposed by Pan and Rickard (2018) and confirms that the involvement of elaborative feedback with practice testing can increase the likelihood of transfer. We believe that when preparing for high-stakes tests, the testing and feedback practice allows students to elaborate the correct-answer and fully encode test questions, as evidenced by the fact that students in testing with feedback group spent more time in high-stakes testing contexts. Although testing itself might lead to better memory for target information according

to the literature (e.g., Hausknecht, et al., 2007; Scharfen et al., 2018), when different items are used, the benefit is diminished.

Another important finding of the current research is that we did not observe significant benefits of testing practice compared with restudy across three experiments. The restudy practice yielded similar effects on post-tests compared to testing with feedback when students were preparing for the high-stakes tests (Experiment 1 and Experiment 3), and the effects were significant compared to the control condition. This runs counter to some previous findings, as extensive prior research identified a significant effect of testing over restudy (Agarwal et al., 2021; Yang et al., 2021). However, almost all prior studies reported a significant testing effect based on the high congruence of learning materials and test materials, using the same test questions in the practice tests during the intervention and the criterion test after intervention (e.g., Greving & Richter, 2018; Wheeler, Ewers, & Buonanno, 2003). According to the transfer appropriate processing framework, memory improves when the cognitive processes engaged during retrieval match the cognitive processes that were engaged when the material was encoded (Morris, Bransford, & Franks, 1977). When the final test emphasises the retention of information that was previously learned, students in the testing group can better match their initial processing to the type of processing required by the final test (Roediger & Karpicke, 2006) and thus outperform the restudy group regarding knowledge retention. In comparison, we used different test items across the pre-test, intervention, and the post-test, as well as various types of test items assessing not only knowledge retention but also understanding and application, as this design is closer to authentic test preparation. Despite the fact that all test items targeted the same general domain, there were still mismatches between the processing at encoding during intervention and the later processing at retrieval during the post-test and the final test. This is a possible reason why the testing effect was not superior to the restudy effect. We can therefore infer

that based on our study that restudy and testing with feedback are equally effective in authentic test preparation contexts where the assessment focuses not only on knowledge retention but also on understanding and application that requires a certain degree of knowledge transfer.

Although students tended to use restudy more frequently in their own learning process (Carpenter et al., 2020), the effectiveness of restudy has not received much research attention in the literature. Instead, it was typically examined as a control condition for the testing effect. A small number of studies have demonstrated the effectiveness of restudy in specific situations, particularly when testing was followed by restudy (Anderson & McDaniel, 2021; McDaniel et al., 2015). For example, Anderson and McDaniel (2021) found that when participants engaged in restudy after taking an initial test, both obtaining test feedback (for all forms of feedback in the study) and having access to the initial test contributed to the success of the restudy. In our research, students took a pre-test on the target concepts before attending an intervention. Participation in the pre-test might boost the restudy effect. McDaniel et al.'s (2015) studies provided some empirical evidence for this assumption. They used a similar design as our experiment to examine the effects of the combinations of multiple testing and restudying practices. In their first experiment, students in the test-restudy-test condition engaged in a multiple-choice quiz first, then restudied learning materials for 8 minutes followed by a second quiz. No feedback was provided after the initial quiz. The results showed that test-restudy- test had a similar effect with the repeated testing, both conditions significantly outperformed the repeated study condition on the final test which was conducted five days after the intervention. In their second experiment, students in the test-restudy-test condition received the graded quiz as the additional material for their restudy, and this produced a significantly more profound effect than repeated testing and repeated studying. Findings of McDaniel et al.'s (2015) experiments suggested that although we did not provide

any feedback directly after the pre-test, the process of engaging in the pre-test might increase the effectiveness of subsequent restudy, possibly because students may monitor their learning through taking the pre-test and regulate their restudy by focusing more on concepts that they were uncertain about during the intervention.

Metacognitive training and practices triggered more testing attempts towards content that participants were not confident in, but did not yield better test performance

We further investigated the assumptions mentioned above on whether the monitoring of learning increase the test preparation effect in Experiment 3. In prior metacognition and self-regulation studies, researchers found that students lack metacognitive awareness during their learning, and this lack of metacognitive awareness is likely because students are rarely explicitly taught about the effective use of strategy (Händel et al., 2020; Pintrich, 2002). To address this problem, we provided students with a short training on the effectiveness of metacognitive strategies, as well as a series of monitoring and control tasks that required them to rate their confidence and make decisions about whether to make more study attempts on the concept. We assumed that students would benefit from monitoring, and control their study attempts effectively, thus achieve a better score.

Findings of this experiment revealed that students showed more practice testing attempts and spent more time on topics for which they were not confident, but no significant association was identified between students' confidence and restudy attempts on learning materials in the restudy plus metacognition group. However, we did not observe clear evidence of the use of metacognitive strategies leading to superior test performance compared to testing and restudy only conditions. Our findings did not provide support for the effective role of metacognitive training and tasks on improving monitoring accuracy and test scores, but we cannot conclude that metacognitive strategies are ineffective. Training and practice time was limited in this experiment, so students' use of metacognitive strategies might not

have been fully developed. The results revealed that the incorporation of metacognition training and monitoring and control tasks has the potential to improve the restudy effect, as demonstrated by the result that the restudy plus metacognition group significantly outperformed the control group on both post-test and the delayed course test, while the restudy only group just did so on the post-test. However, based on findings from Experiment 3, it is difficult to determine the mechanism of this effect. Students may have had made better use of metacognition strategies during their own preparation for the course test, but this study yields limited information about how extensively students self-prepare for the course test. Despite the fact that the testing plus metacognition did not yield larger score gains than testing alone, this finding does not dismiss the benefits of metacognitive monitoring and control in testing conditions. It is highly likely that feedback provided after testing questions stimulates students' monitoring of their study, thus adding metacognitive practices may have had no additional effects. Educational researchers have written extensively about the role of feedback for learning (e.g., Black & Wiliam, 1998; Butler & Winne, 1995; Hattie & Timperley, 2007; etc.) and it can enhance the effects of retrieval practices (Butler et al., 2007). Researchers believed that as feedback provides students with information about their learning status, it allows them to reconsider the learning task and their engagement in subsequential learning sessions (Barber et al., 2011). Thus, in testing conditions, students can obtain cognitive benefits from retrieval attempts, and further engage in metacognitive monitoring of their knowledge based on the provision of feedback. Students' motivation to keep on studying might also increase, potentially by making one's own progress in learning more transparent and enhancing experience of competence (Abel & Bäuml, 2020).

No significant test preparation effect is found in the low-stakes testing context

In contrast to findings gathered from high-stakes testing contexts (Experiment 1 and 3), none of the three experimental conditions (testing with feedback, testing without feedback, and

restudy) produced significant advantages over the no preparation condition on the low-stakes test results (Experiment 2). This is a new finding adding to the literature, because to our knowledge, there was no existing study comparing test preparation effects on high-stakes test results versus low-stakes test results. Hinze and Rapp (2014) examined the effects of performance pressure with either high-stakes or low-stakes quizzes on memory and comprehension and found that the testing effect was attenuated by high performance pressure. They manipulated performance pressure by whether linking participants' performance with money incentives (high-stakes quiz groups) or not (low-stakes quiz groups). Though our study also manipulated a high-stakes test preparation condition (Experiment 1) and a low-stakes test preparation condition (Experiment 2), there may be important differences between our study and Hinze and Rapp's that led to different results. In their study, the high-stakes group reported higher state anxiety when they practiced quiz questions. In our experiment, students might be more motivated to learn before a high-stakes test rather than getting more anxious. A recent study has shown that hope of success moderates the effect of testing (Heitmann, et al., 2022). Since we did not collect affective data such as motivation, test anxiety, hope to success, and learning attitudes, we are unable to confirm whether the difference between test preparation effects in high-stakes versus low-stakes contexts relate to students' affective characteristics, such as motivation to learn or hope of success. These can be explored in future studies.

Limitations

There are two important limitations regarding this research's experimental design. First, we regard restudy as an integral strategy and did not break it down. The restudy conditions' internal validity might have been compromised as students might use different techniques, such as re-reading texts or rehearsal, to varying degrees. Second, the interpretation of the

course test results might be problematic. Considering that this research was conducted in an authentic educational setting, students in Experiments 1 and 3 prepared for real high-stakes tests when they participated in the study. We were unable to control for their exposure to relevant course materials after the intervention. Thus, the effect of the intervention on the final course test results may be compromised, and students' performance is likely to have been affected by how well they prepared on their own after the intervention. Nonetheless, this is a feature of authentic test preparation.

Conclusions

In conclusion, this research indicates that testing practice with feedback and restudy are equally effective strategies to use in a cram session right before the high-stakes test, where the assessment focuses not only on knowledge retention but also on understanding and application that requires a certain degree of knowledge transfer. Testing with feedback could enhance long-term learning, whereas the effect of restudy might diminish after a period. The combination of metacognitive training and tasks with restudy practice was found to be effective for long-term learning, indicating that the metacognitive monitoring and control might boost the success of restudy practice. Though metacognitive training and tasks did not add to the effects of testing with feedback, they drove students to devote more practice testing effort to unconfident topics. These findings should encourage future interventions to support students in developing metacognitive strategies altogether with the use of testing and restudy practice in authentic test preparation contexts, and further explore how the metacognitive training and practice influence students' regulation of their restudy. In addition, this study contributed to the understanding of how the stakes of tests affects students' test preparation and learning. In the low-stakes context, testing and restudy did not produce significant score gains compared to the no preparation condition. A next step towards this finding consists of

exploring whether and how stakes of tests affect students' motivation and learning effort, and how learning attitudes and effort influence test preparation. Furthermore, as this research was conducted in China, it is possible that students' attitudes towards test preparation and the two strategies may differ from those in the West, thus play a role in the different findings observed in this research. It is important to take this context into consideration when interpreting findings of this research. Further comparative investigation can further explore the potential cultural differences.

References

Abel, M., & Bäuml, K.-H. T. (2020). Would you like to learn more? Retrieval practice plus feedback can increase motivation to keep on studying. *Cognition*, *201*, 104316. https://doi.org/10.1016/j.cognition.2020.104316

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

Anderson, F. T., & McDaniel, M. A. (2021). Restudying With the Quiz in Hand: When Correct-Answer Feedback is No Better Than Minimal Feedback. *Journal of Applied Research in Memory and Cognition*, *10*(2), 278–288. https://doi.org/10.1016/j.jarmac.2020.10.004

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, *61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Barber, L. K., Bagsby, P. G., Grawitch, M. J., & Buerck, J. P. (2011). Facilitating Self-Regulated Learning With Technology: Evidence for Student Motivation and Exam

Improvement. *Teaching of Psychology*, *38*(4), 303–308. https://doi.org/10.1177/0098628311421337

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS: 88. Chance, 14(1), 10–18.

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Carpenter, S. K., Endres, T., & Hui, L. (2020). Students' Use of Retrieval in Self-Regulated Learning: Implications for Monitoring and Regulating Effortful Learning Experiences. *Educational Psychology Review*, *32*(4), 1029–1054. https://doi.org/10.1007/s10648-020-09562-w

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A Classroom Study on the Relationship Between Student Achievement and Retrieval-Enhanced Learning. *Educational Psychology Review*,*28*(2), 353–375. https://doi.org/10.1007/s10648-015-9311-9

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. https://doi.org/10.3758/BF03202713

Cogliano, M., Kardash, C. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology*, 56, 117-129.

Freda, N. M., & Lipp, M. J. (2016). Test-Enhanced Learning in Competence-Based Predoctoral Orthodontics: A Four-Year Study. *Journal of Dental Education*, *80*(3), 348–354. https://doi.org/10.1002/j.0022-0337.2016.80.3.tb06091.x

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology*, *60*(7), 991–1004. https://doi.org/10.1080/17470210600823595

Goossens, N., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H., & Zwaan, R. (n.d.). *The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning*. https://doi.org/10.1016/j.jarmac.2014.05.003

Greving, S., & Richter, T. (2018). Examining the Testing Effect in University Teaching: Retrievability and Question Format Matter. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02412

Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction*, *65*, 101245. https://doi.org/10.1016/j.learninstruc.2019.101245

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heitmann, S., Grund, A., Fries, S., Berthold, K., & Roelle, J. (2022). The quizzing effect depends on hope of success and can be optimized by cognitive load-based adaptation. *Learning and Instruction*, *77*, 101526. https://doi.org/10.1016/j.learninstruc.2021.101526

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (Sometimes) Enhances Learning: Performance Pressure Reduces the Benefits of Retrieval Practice. *Applied Cognitive Psychology*, *28*(4), 597–606. https://doi.org/10.1002/acp.3032

Janes, J. L., Dunlosky, J., & Rawson, K. A. (2018). How Do Students Use Self-Testing Across Multiple Study Sessions When Preparing for a High-Stakes Exam? *Journal of Applied Research in Memory and Cognition*, *7*(2), 230–240. https://doi.org/10.1016/j.jarmac.2017.11.003

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D. (2017). Retrieval-Based Learning: A Decade of Progress. In *Learning and Memory: A Comprehensive Reference* (pp. 487–514). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(4), 989.

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of Practice on Aptitude and Achievement Test Scores. *American Educational Research Journal*, *21*(2), 435–447. https://doi.org/10.3102/00028312021002435

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition*, *3*(3), 171–176. https://doi.org/10.1016/j.jarmac.2014.04.002

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied*, *21*(4), 370–382. https://doi.org/10.1037/xap0000063

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Peterson, D. J., & Wissman, K. T. (2018). The testing effect and analogical problem-solving. *Memory*, *26*(10), 1460–1466. https://doi.org/10.1080/09658211.2018.1491603

Pintrich, P. R. (2002). The Role of Metacognitive Knowledge in Learning, Teaching, and Assessing. *Theory Into Practice*, *41*(4), 219–225. https://doi.org/10.1207/s15430421tip4104_3

Rawson, K. A., & Dunlosky, J. (2012). When Is Practice Testing Most Effective for Improving the Durability and Efficiency of Student Learning? *Educational Psychology Review*, *24*(3), 419–435. https://doi.org/10.1007/s10648-012-9203-1

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The Power of Successive Relearning: Improving Performance on Course Exams and Long-Term Retention.

Educational Psychology Review, *25*(4), 523–548. https://doi.org/10.1007/s10648-013-9240-4

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing Effect in the Psychology Classroom: A Meta-Analytic Perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096-1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1)*, 24-31.

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784–802. https://doi.org/10.1080/09658211.2013.831454

Thomas, R. C., Weywadt, C. R., Anderson, J. L., Martinez-Papponi, B., & McDaniel, M. A. (2018). Testing Encourages Transfer Between Factual and Application Questions in an Online Learning Environment. *Journal of Applied Research in Memory and Cognition*, *7*(2), 252–260. https://doi.org/10.1016/j.jarmac.2018.03.007

van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology*, *36*(3), 212–218. https://doi.org/10.1016/j.cedpsych.2010.10.004

Wheeler, M. A., & Roediger, H. L. (1992). Disparate Effects of Repeated Testing: Reconciling Ballard's (1913) and Bartlett's (1932) Results. *Psychological Science*, *3*(4), 240–246. https://doi.org/10.1111/j.1467-9280.1992.tb00036.x

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571–580. https://doi.org/10.1080/09658210244000414

Author Contributions

Zhanxin Hao: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing-Original Draft.

Jo-Anne Baird: Conceptualization, Writing-Reviewing & Editing, Supervision

Yasmine El Masri: Conceptualization, Writing-Reviewing & Editing, Supervision

Chapter 6. Paper 3

Chapter 6 presents the third research paper (Paper 3), which probes how participants from a teacher training school in China regulated their combined uses of testing and restudy, factors associated with their strategy choices, as well as their perceptions on the efficacy of the two strategies. Note that in this paper, the term 'rereading' is used to refer to the strategy of reading statements on educational psychology topics, which is referred to as 'restudy' in Paper 2 and other chapters in this thesis. The use of 'rereading' is intended to avoid confusion for readers. Since students in this study were allowed to review a topic multiple times and the study analyses their attempts at repeated engagement with each topic, readers might interpret these repeated engagements as 'restudy', thus the term 'rereading' is used throughout the paper to avoid confusion. Results indicate that students in this study had an obvious preference to testing, and they seemed to be able to recognise both the direct benefits of testing and the indirect benefits of testing, i.e., testing as a monitoring tool to guide the direction of subsequent learning. Students' preference and trust in testing have been discussed from a sociocultural perspective in this paper.

Chinese students' use of testing and rereading strategies for test preparation

Zhanxin Hao[1]; Jo-Anne Baird[1]; Yasmine El Masri[2]

[1] Department of Education, University of Oxford, Oxford, UK

[2] Ofqual, Coventry, UK

Corresponding details: zhanxin.hao@education.ox.ac.uk

Funding: This work was supported by Chinese Scholarship Council [grant number: 201808060181].

Chinese students' use of testing and rereading strategies for test preparation

Abstract

As high-stakes testing has been increasingly used in educational contexts, students sometimes prepare for the test independently. It is crucial to depict students' self-preparation processes by which they use different learning strategies during test preparation and investigate the reasons behind students' choices, to enable them to regulate their test preparation in more efficient ways. This study was conducted in a teacher training college in China, and provided students with a series of educational psychology topics to review before their course test. Students were allowed to choose either testing or rereading to review each topic as many times as they wanted, and they rated the difficulty of each topic, as well as their confidence in knowledge mastery of each topic before and after every study attempt. Results suggested that students used one-time testing most often and were reluctant to repeatedly reread a topic. Both results of general linear mixed effect models and qualitative data on students' perceptions of the two strategies indicated that students' strategy choice was predicted by their confidence and the type of question used for testing. This research suggested that Chinese students in this study could recognise both the direct and indirect effect of testing, which might be culturally distinctive.

Keywords: test preparation; Chinese students; testing; rereading; study strategies; monitoring and control; students' confidence

Chinese students' use of testing and rereading strategies for test preparation

1. Introduction

Since high-stakes educational tests are widely used in educational systems for various purposes such as selection, qualification, student monitoring, etc. (Newton, 2007), test preparation has become a common activity in the learning process. Although students can take test preparation courses at schools and commercial institutes, they also need to rely on themselves to regulate their test preparation. Self-regulation in test preparation processes is essential; as students progress to higher grades, they need to study independently and be more accountable for test results. For instance, in contrast to the compulsory educational stage where teachers and schools are also held accountable for students' test performance, when it comes to higher education contexts, most of the learning is unsupervised, and students are the important stakeholder in such high-stakes educational tests (e.g., course tests in college, or language tests used for undergraduate or post-graduate admission). Therefore, controlling and using suitable learning strategies is critical to effective test preparation, as suggested by self-regulated learning theory (Bjork et al., 2013; Dunlosky & Metcalfe, 2009; Winne & Hadwin, 1998; Zimmerman, 2000).

Previous surveys reported that students engage in a variety of learning activities and employ a range of strategies to prepare for tests, such as repeated reading, practice testing, note-taking, concept mapping, and so on (Karpicke et al., 2009). These strategies are not equally effective, and unfortunately, learners do not always make effective choices (Blasiman et al., 2017; Karpicke et al., 2009; Yan et al., 2014). We also need to be aware that learning is a dynamic process in which students usually blend various learning strategies. Therefore, it is crucial to depict the processes by which students use different learning strategies during test

preparation and investigate the reasons behind students' choices, in order to enable them to regulate their study in more efficient ways.

1.1 How do students regulate the strategy decisions for test preparation?

Some previous studies surveyed how students study on their own for an upcoming test, revealing that repeatedly reading the textbook (i.e., re-reading) and taking practice tests (i.e., testing) are the two most common strategies (Blasiman et al., 2017; Karpicke, 2009). Some earlier studies found that students relied most on repeatedly reading the textbook over practice testing or using flashcards (Blasiman et al., 2017; Hora & Oleson, 2017; Karpicke, 2009; Morehead et al., 2016). Recent studies demonstrated possible patterns in students' strategy choices. Researchers found that students tend to test themselves on easy materials or those they have recently studied and restudy materials that are perceived as difficult or which have not recently studied (Karpicke, 2009; Toppino et al., 2018; Tullis et al., 2018). Most experimental studies involved a fixed amount of practice and allowed students to choose a certain strategy over single sessions. However, in authentic educational contexts, it is unlikely that students engage in a single acquisition session, using one strategy solely during test preparation, as in experimental studies. Badali et al. (2022) allowed students to engage in as many learning trials as they wanted, and they found that students tended to restudy difficult items at the initial stage and shift to using practice tests when they were allowed to engage in multiple learning trails.

Rivers (2021) concluded that students choose testing mainly when they expected that retrieval would be successful. She identified four scenarios in which students tended to use testing: 1) when the material is perceived as easy to recall; 2) when the lag between studying and testing is short; 3) when they receive feedback on their performance; and 4) later on in the learning process. These patterns, identified by previous research and summarised by Rivers, possibly reveal that students were using testing as a monitoring tool, which helps

them to diagnose their knowledge mastery and inform decisions about whether further study is needed.

1.2 How do students perceive testing and rereading?

The strategies students rely on most are not necessarily the most effective ones. Why do students not use testing more often in their test preparation? Students' study decisions are usually driven by which strategies learners perceive to be effective (Kirk-Johnson et al., 2019), thus one possible explanation is that students fail to recognise the benefits of testing, while some 'less effective' strategies, such as rereading, were valued by students (Anthenien et al., 2018; Blasiman et al., 2017; Morehead et al., 2016; Rivers, 2021). However, some recent research revealed that though students do not always understand the direct benefit of testing for memory, many students do seem to appreciate its indirect benefits for monitoring, as evidenced by some recent findings that students tended to use testing in latter stages of learning (Janes et al., 2018; Karpicke, 2009; Kuhbandner & Emmerdinger, 2019). For example, Janes et al. (2018) recorded students' practice in four sessions of the course, which began from the instructor's introductory session and ended prior to the course exam. They found that students used rereading more than testing in earlier sessions, while they tended to switch to testing in later sessions when they learned the content better. Findings from these studies suggest that students may regard testing as a way to diagnose their learning, but they might lack metacognitive awareness of testing as a tool to enhance memory and learning (Rivers, 2021).

Another possibility as to why students use the testing strategy infrequently is that students might regard testing as an effortful strategy and are reluctant to use it. Krik-Johnson et al. (2019) argued that students' perceptions about strategy effectiveness can be misguided by how much mental effort is experienced when using a particular strategy. Supportive evidence was found across multiple laboratory experiments in their study: the more students

perceived a study strategy as mentally effortful, the less they judged it to be effective for learning; this in turn predicted their choices of study strategies. This hypothesis seems to be reasonable, as some studies found that students tended to mass their practice close to the exam, and their actual strategy use diverged from the intentions they initially reported (Blasiman et al., 2017). In another word, students may want to choose some easier-to-use strategies when they are running out of time.

1.3 The current study

Self-regulated learning theory holds that learners may adjust their usage of strategies according to their monitoring of learning (Winne & Hadwin, 1998; Zimmerman, 2000). Based on this theory and findings of previous studies, here we build and evaluate a hypothetical model of students' adaptive self-regulated test preparation based on both cognitive and metacognitive science: Students' strategy decisions for test preparation are determined by both their predictions of their performance and their perceptions of the efficacy of each test preparation strategy. Students often do not have direct access to their objective level of knowledge mastery during study. Their predictions of how well they will perform on a topic in the upcoming test may be linked to not only their subjective confidence in knowledge mastery but also the perceived difficulty of the specific topic and the relevant assessment task. Students' predictions might be dynamically adjusted as students learn more and progress in their learning. This may consequently influence students' adaptive use of learning strategies, which might ultimately determine how well they perform on the test. Many links in this chain of influences have been suggested by some existing empirical evidence, such as how students' perceptions of strategies affect their actual strategy choice (Blasiman et al., 2017; Kirk-Johnson et al., 2019), how students' judgements of learning relate to their strategy decisions (Karpicke, 2009; Toppino et al., 2018), and the relationship between strategy use and academic achievement (Hartwig & Dunlosky, 2012). However, to

our knowledge, no existing study has simultaneously assessed all of these links, especially how students' perceptions of the effectiveness of strategies and predictions of their performance associate with the adaptive use of learning strategies (i.e., the repetitive use of a certain strategy, or the blended use of different strategies) within the same participants during their self-regulated test preparation processes.

In this study, we directly test this model by investigating students' use of test preparation strategies in an authentic test preparation context and exploring what factors influence strategy use. Specifically, to prepare for an upcoming educational psychology test, our participants used an online platform where their strategy choices, monitoring of study (i.e., rating of the confidence and difficulty of each topic), and control of study (i.e., whether to study the topic repeatedly and use what strategy) were recorded and the relationship between the monitoring and control was analysed. To recap, it is generally accepted in the literature that students' perceptions of strategy effectiveness affect their decision-making, even though some researchers believed that students may not be aware of the direct effect of testing on learning, while others indicated that students may view testing as an effortful approach and thus ineffective. These inferences were mainly drawn from observations of students' study behaviours, which needs to be confirmed by asking students' views on the effects of testing and rereading. In the current study, we used a survey with several open-ended questions to explore students' perceptions of the effectiveness of test preparation strategies (i.e., testing and rereading) and conditions under which they would use each strategy. This would allow us to develop a better understanding of students' strategy choices and the possible mechanisms underlying their choices. Based on the findings of this study, we hope to portray a fine-grained picture of the what, when and why of students' study decisions for test preparation.

The following research questions are explored in this study:

1) How did students use the rereading and testing strategy when they prepared for a high-stakes educational psychology test?

2) What factors affected students' strategy decisions?

3) How did students perceive the use and effects of testing and rereading strategies on test preparation?

2. Method

2.1 Participants

Participants were third-year students from a pre-service teacher training college in China. This college provides teacher development programmes for early years and primary school education. Students enrolled in these programmes are required to pass the National Teacher Qualification Examination during their study and the majority of graduates work in kindergartens or primary schools in rural areas in China. One of the subtests of the National Teacher Qualification Examination is Educational Psychology, which is also one of the compulsory courses in the school. The research was conducted during educational psychology classes, providing students with an online test preparation session for the upcoming course test. Students from three classes were invited to participate in the research in exchange for course credit (N = 152, 65% female, *Mean age* = 17.51). All three classes included in the research were taught by the same teacher.

2.2 Design

This study was a quasi-experimental study, using a non-equivalent pretest-posttest design. It was conducted in an authentic educational context in which students were preparing for a teaching qualification and sitting a summative test at the end of their course. We provided an online preparation session for students from three classes where they learned 16 educational psychology topics that would be assessed in the summative test. Students took the pre-test

and post-test, and their performance on the tests was compared with students in business-as-usual conditions to indicate the effectiveness of preparation. However, as this article is primarily concerned with understanding students' study decisions, the effectiveness of these decisions is beyond the scope of this article and is not analysed and discussed here.

During the online preparation session, students regulated their learning by deciding whether to study the topic for multiple times, selecting one of the two strategies, testing or rereading. Students were required to do the selection for each study attempt. All their study decisions were recorded. We also investigated whether students' monitoring of learning was associated with their study decisions by asking students to rate the difficulty of each topic and rate their confidence in the knowledge mastery before each study attempt. Since students' confidence in knowledge might change as learning progressed, they were asked to rate their confidence before and after each study attempt, so as to illustrate the change of their confidence in mastering the topic after each study attempt.

After the online session, students' attitudes towards the usage of the two strategies under high-stakes testing conditions were collected using several open-ended questions. To be specific, students were asked which strategy they believed was more effective and why, as well as in what conditions they would use the two strategies.

2.3 Materials

Test preparation materials

Materials used in the online test preparation session were from the educational psychology course taken by the students in their third year. This study was carried out at the end of the first term of the course, when students had learned about 40% of the whole educational psychology curriculum, including four chapters: 'Psychological development and education of primary school students', 'Learning and learning theories', 'Learning strategies', and

'Knowledge and skills'. A total of 16 topics, four from each chapter, were sampled for the test preparation session. For each topic, one statement (for rereading) and one testing question (for testing) were created by summarising the key information about the topic based on the textbook or the instructor's PowerPoint slides used for teaching. The researcher and the course teacher discussed each statement and each testing question to make sure they were stated properly and accurately. Examples of practice questions and statements can be found in Appendix A. The intervention materials were presented to participants using the PsyToolkit platform (Stoet, G. 2010, 2017).

Monitoring and control tasks

Based on the literature, we were interested to explore whether and to what extent students' monitoring of their learning influenced their study decisions. During the intervention, students in the intervention group did a series of metacognitive monitoring and control tasks, including rating the difficulty of the topic, rating confidence before and after learning each topic, and making decisions on repeated studies. To be specific, students were asked to rate the difficulty of each topic on a 10-point Likert-type scale ranging from 1 (very easy) to 10 (very difficult). Note that in order to separate students' judgements of their confidence for each topic and the difficulty level of each topic they perceived, we reminded students before the intervention that when they rated the difficulty of each topic, they should try to be objective, comparing the topic with others in the curriculum and giving an estimated rate based on their perceptions. Before and after reviewing the topic (no matter what strategy they used for the review), students were also required to indicate their confidence of knowledge mastery from 1 (very unsure I have mastered the topic) to 10 (very sure I have mastered the concept).

2.4 Procedure

Instructions & Training

Before the intervention, a training session was provided to all participants to introduce the research design and procedures, and to instruct them how to use the online test preparation platform. Students practiced with several example questions to familiarise themselves with the platform and tasks. After the training, a question-and-answer session was held to solve students' further questions regarding the experiment.

Test preparation session

During the session, students engaged in 60-minute self-paced learning through the PsyToolkit platform (Stoet, 2010, 2017). They could access it via laptops, tablets or mobile phones. Students were encouraged to use whichever device they were most comfortable with, and they were also allowed to use paper and pencil to facilitate their practice (for example, taking notes). The practice was self-paced, and the researcher reminded students of the remaining time every ten minutes.

All materials were provided on the PsyToolkit platform. Each topic of the to-be-learned concept was presented on the screen first, after which students were required to rate the difficulty of the topic and their confidence of knowledge mastery. Then they made a decision about which strategy to use to learn this concept, choosing from rereading the statement or answering a practice question. The type of the practice question (multiple-choice or short response) was provided to help students make the decision.

After reviewing each topic, students were required to rate their confidence of mastery of this topic. Then, they were asked to decide whether to review the same topic again or move to the next topic. Students made as many attempts as they wanted on each topic, using one of the two strategies (i.e., testing and rereading) in each attempt. A series of screenshots

(translated from Chinese) could be found in Appendix B that illustrates tasks included in each learning attempt during the online test preparation session.

After the intervention, students were asked to write down their perceptions of the two strategies by answering questions 'What do you think of the rereading strategy, when do you want to use it?', 'What do you think of the testing strategy, when do you want to use it?', 'Which strategy do you prefer?', 'Which one is more effective, in which conditions?'. One week after the experiment, a 90-minute final course exam was administered.

2.5 Analytical strategy

2.5.1 Descriptive analysis

This study aimed to depict students' strategy uses during test preparation, what factors influence their strategy uses, and how do students understand different strategies. We first conducted descriptive analysis using SPSS 28. The total number of study attempts by all students on the 16 topics was tallied, which was further broken down into the numbers of attempts using testing and rereading. For each of the three attempts, the frequencies of choosing testing and rereading were counted. Besides, the average ratings of the difficulty on the 16 topics were analysed, so were the average ratings of confidence students made before and after learning each topic.

2.5.2 Generalised linear mixed-effects models (GLMMs)

To obtain an overall picture of how both students' metacognition and topic characteristics as a composite model explained variance in students' strategy selection, we conducted two-level generalised linear mixed-effects models (GLMMs) using the *glmer* function in the 'lmer4' package (Version 1.1–21) (Bates et al., 2015) in the R Studio (R Core Team, 2020). Three separate models were conducted to analyse what factors could predict students' strategy

choices in their first and second study attempts, as well as what influenced their repeated study decisions[1].

Regarding the first model, the dependent variable was students' strategy selection for the first study attempt, which was a binary variable and was dummy coded (the testing strategy was coded as 0 and the rereading strategy was coded as 1). The fixed effects included metacognitive ratings (confidence and perceived difficulty of the topic) and type of the testing question (treated as a nominal factor consisting of two values: multiple-choice question and short-response question). Without the inclusion of random effects, a fixed-effects-only model would be a standard generalised linear regression which would only provide predictions for the participants and/or all items as a whole group and does not accommodate individual variations around the group's average tendency on strategy choices. A random intercept of student and a random intercept of item, therefore, were specified to allow for individual variations in each student on each item. According to Barr et al. (2013), the key to determining random effect structure is figuring out which factors are within-students or within-topics, and where any pseudoreplications are located in the design. We specified a random slope of student confidence and perceived difficulty to allow for variation in the strategy selections of these topics. In addition, a random slope for the confidence, difficulty and question type was also specified to allow for individual variation of students. The following model was used to examine the source of variances among students' strategy choice.

$$\begin{aligned} &Strategy\ choice\ (first\ attempt) \\ &\quad = Intercepts + InitialConfidence + Perceived\ Difficulty \\ &\quad + Question\ Type + Random\ Effects\ of\ Students \\ &\quad + Random\ Effects\ of\ topics \end{aligned}$$

[1] We tried to run a longitudinal model using students' study attempts as the longitudinal variable. However, this model failed to converge. Therefore, we chose to perform three separate models.

The second model examined what factors were associated with students' decisions on whether to repeat their study on the topic. Students repeated study decision was dummy coded as 0 (students did not study for the second time) and 1 (students studied for the second time), which was used as the dependent variable. The fixed effects of this model included three predictors: 1) strategy of first choice; 2) confidence after first attempt; 3) perceived difficulty of the topic; 4) type of the testing question. The random slope for confidence and perceived difficulty on topic level and for confidence, difficulty and question type on student level were specified.

$$
\begin{aligned}
&Repeated\ study\ decision \\
&\quad = Intercepts\ + Strategy\ choice\ of\ first\ attempt \\
&\quad + Confidence\ after\ first\ attempt + Perceived\ Difficulty \\
&\quad + Question\ Type + Random\ Effects\ of\ Students \\
&\quad + Random\ Effects\ of\ topics
\end{aligned}
$$

In the third model, students' strategy choices in their second study attempts were used as the dependent variable. The fixed effects included four predictor variables: 1) strategy choice of first attempt; 2) perceived difficulty of the topic; 3) confidence of the topic after first attempt; 4) type of the testing question. Similar with the first model, a random intercept of student and a random intercept of item were specified to allow for individual variations in each student on each item. A random slope of confidence and perceived difficulty on topic level and a random slope of confidence, difficulty and question type on student level were specified.

$$
\begin{aligned}
&Strategy\ choice\ (Second\ attempt) \\
&\quad = Intercepts\ + Strategy\ choice\ of\ first\ attempt \\
&\quad + \ Confidence\ after\ first\ attempt + Perceived\ Difficulty \\
&\quad + Question\ Type + Random\ Effects\ of\ Students \\
&\quad + Random\ Effects\ of\ topics
\end{aligned}
$$

Before performing each GLMM model described above, ANOVA models were performed to compare the model fit between the null model with random intercept and the null model with both random intercept and random slope. The results showed that for each model, the random slope model had significantly better model fit. The model fit information can be found in Appendix C.

2.5.3 Qualitative analysis

Students' perceptions of the conditions under which they would use testing and rereading were analysed by thematic analysis. Braun and Clarke (2006) defined thematic analysis as 'a method for identifying, analyzing and reporting patterns within data' (p. 79). They provided a guideline for thematic analysis, including: a) familiarising with data, b) generating initial codes, c) searching for themes, d) reviewing themes, e) defining and naming themes, and g) producing the report (p. 88). This six-phase framework was followed in the analysis of the qualitative data.

The course teacher was invited to assist the first author of the article in coding qualitative data. The course teacher had a degree in Bachelor of Education and was familiar with the procedures of thematic analysis. Both the course teacher and the first author were native Chinese speakers; therefore, they coded the transcripts in Chinese during the coding stage. In the beginning of the coding, the first researcher and the course teacher coded the responses of 30 students (20%) separately, and they agreed with each other for over 85% of the shared coding data. Disagreements were reconciled by a discussion. Then, the first author

of the paper read all remaining students' responses, applying the codes to label data extracts. The first author identified the recurrent codes and possible themes during coding. The defined themes were presented to the course teacher and were discussed. The finalised themes and codes were translated into English by the first author.

3. Results

3.1 Students' test preparation behaviours and metacognition

3.1.1 Test preparation behaviours

During the test preparation session, students were allowed to make as many study attempts as they needed on each item, using one of the two strategies (i.e., testing or rereading) in each attempt. In total, 152 participants made 3643 study attempts for the 16 topics, with 33.24% of the study attempts (N = 1211) being repeated studies (i.e., the second, third, and fourth study attempts). As shown in Table 16, for the first study attempt, the use of testing (N = 1181) is weighted similarly to the use of rereading (N = 1251), while for the second and third study attempts, testing was used more frequently than rereading. The Chi-square test showed that there was significant association between study attempt and strategy choice, X^2 (3, N = 3643) = 262.03, p < .001.

Table 16 Crosstabulation of Study Attempt and Strategy Use.

Study attempt	Strategy		Total
	Testing	Rereading	
First	1181	1251	2432
Second	808	237	1045
Third	109	41	150
Fourth	14	2	16
Total	2112	1531	3643

In most cases, participants studied a topic only once or twice, and there were only 4.6% (*N*=166) of instances where students studied the topic for the third and fourth times. Therefore, it is worth delving deeply into how students' strategy choices changed during the first two rounds of study to identify possible patterns. As shown in Figure 12, about 70% (*N* = 827) of the testing trials in the first study attempts were not followed by further study attempts, about 20% (*N* = 232) of the testing trials in the first attempts were followed by a second testing trial, and the remaining 10% (*N* = 122) of the testing trials were followed by a rereading trial. As for the strategy of rereading, about 46% (*N* = 577) of the rereading trials in the first attempts were followed by a testing trial in the second attempt, while only 9% (*N* = 114) of them were followed by another rereading trial. The remaining 54% (*N* = 560) of rereading trials in the first attempts were the only attempt.

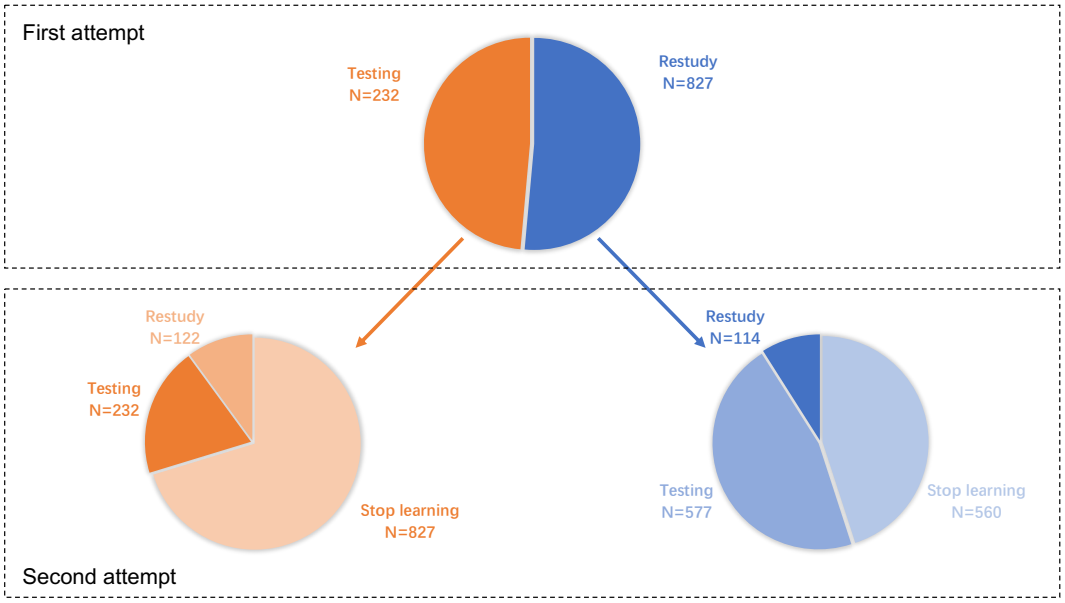


Figure 12 Participants' strategy choices among their first two study attempts.

Most participants used both testing and rereading interchangeably, but there were five out of the 152 participants (about 3%) used rereading as the only strategy across their study

attempts. No participant adopted testing as the only strategy, but most participants (66%, N = 100) used testing more often than rereading.

To further understand how students blended the use of the two strategies to prepare for the test, we coded their strategy use for each topic into the following five patterns: 1) solely using testing to study a topic for just one time; 2) solely using rereading to study a topic for just one time; 3) repeatedly using testing to study a topic; 4) repeatedly using rereading to study a topic; 5) using testing and rereading in turn. Frequency of these five patterns were presented in Table 17. It was found that solely using testing to study a topic was the most frequent strategy (34%), while repeated rereading of a topic was the least frequent strategy (4%).

Table 17 The frequency of different strategy use patterns.

Pattern of strategy use	Frequency	Percentage
1) Using Rereading to study a topic for one time	560	23%
2) Using testing to study a topic for one time	827	34%
3) Repeatedly using rereading	93	4%
4) Repeatedly using testing	221	9%
5) Using testing and rereading in turn	731	30%

3.1.2 Students' metacognition

In the current study, students were asked to rate the difficulty and their confidence of knowledge mastery for each topic on 1-10 scales before they reviewed the material. They were also asked to rate their confidence of knowledge mastery for each topic after every study attempt. Students' metacognitive ratings are presented in Table 18.

Table 18 Students' average metacognitive ratings on each topic.

Topic	N	Perceived difficulty	Pre-Confidence (before the first attempt)	Post-Confidence (after the last attempt)
1	152	5.61 (2.35)	5.13 (2.55)	7.39 (2.28)
2	152	5.84 (2.41)	5.47 (2.65)	7.17 (2.27)
3	152	5.90 (2.38)	5.38 (2.51)	7.66 (2.08)
4	152	6.61 (2.46)	4.39 (2.51)	7.36 (2.32)
5	152	5.95 (2.38)	5.43 (2.55)	8.27 (1.96)
6	152	5.87 (2.44)	5.71 (2.55)	7.01 (2.35)
7	152	5.86 (2.38)	5.46 (2.39)	7.53 (2.12)
8	152	7.01 (2.24)	4.07 (2.24)	7.36 (2.23)
9	152	6.34 (2.30)	5.13 (2.54)	7.82 (2.10)
10	152	6.76 (2.41)	4.98 (2.53)	7.43 (2.29)
11	152	6.95 (2.26)	4.10 (2.32)	6.63 (2.25)
12	152	6.64 (2.15)	4.49 (2.22)	6.34 (2.24)
13	152	6.01 (2.21)	5.09 (2.37)	8.11 (1.94)
14	152	6.47 (2.11)	4.73 (2.13)	5.91 (2.22)
15	152	6.01 (2.38)	5.31 (2.59)	7.84 (1.82)
16	152	6.21 (2.25)	5.09 (2.25)	5.80 (2.36)

The relationship between students' confidence ratings was analysed through correlation analysis. The results indicated that students' perceived difficulty negatively correlated with their pre-confidence (r = -0.34, p < .001) and post-confidence (r = -0.08, p < .001), suggesting that students had less confidence for more difficult topics. In addition, students' pre-confidence and post-confidence were positively correlated (r = 0.26, p < .001).

Students' ratings of confidence and perceived difficulty on each topic were not associated with question type. The independent sample t-tests suggested that there was no significant difference of perceived difficulty between topics tested by multiple-choice

questions and those tested by short-response questions (t = -0.28, p = .779). Similarly, no significant difference in students' confidence among topics tested by two question types was identified (t = -0.26, p = .793).



Figure 13 Patterns of students' strategy uses and metacognitive ratings.

To understand whether there were differences in students' metacognitive ratings between different strategy use patterns, one-way ANOVA models were performed. The results demonstrated that there was a significant difference in the perceived difficulty between different strategy patterns (F = 6.55, p < .001, η_p^2 = .011). Post hoc analysis showed that the mixed use of rereading and testing was associated with significantly higher ratings of perceived difficulty than the strategies only involving testing (testing once and repeated testing) (ps < .001), while rereading once was associated with significantly higher difficulty rating than repeated testing (p < .05).

Similarly, the difference in pre-confidence ratings between the five patterns was also significant ($F = 5.32$, $p < .001$, $\eta_p^2 = .009$). The mixed use of rereading and testing was associated with significantly lower mean ratings of pre-confidence than the testing only strategies (testing once ($p < .01$) and repeated testing ($p < .001$). Using pre-confidence as the covariate and post-confidence as the dependent variable, we further ran a one-way ANCOVA to explore the difference on students' confidence after learning by different strategies. The results revealed a significant group difference ($F = 36.80$, $p < .001$, $\eta_p^2 = .057$). Post hoc analyses showed that the use of repeated rereading was associated with the lowest average confidence gain after studying compared to the other four patterns (ps < .001 to testing once, mixed use of testing and rereading, and repeated testing; when compared to rereading once, p < .05). The pattern of rereading once associated with the second smallest mean post-confidence gain, while testing once, mixed use of testing and rereading, and repeated testing were associated with larger post-confidence rating gains compared to rereading once (ps < .001). To sum up here, the results indicated that testing practice associated with relative higher confidence levels, while rereading practice associated with relative lower confidence levels.

3.2 Relationship between students' study behaviours and metacognition

The main objective of this research was to understand the extent to which students' study behaviours were affected by their metacognition. The analysis in the above section regarded all participants and all topics as a whole, and did not accommodate individual variations. Three GLMMs that allowed variations of individual students and individual topics were run to investigate what factors predicted students' first strategy choice, repeated study decision, and second strategy choice.

3.2.1 Model 1: Students' strategy choice on first study attempt

We performed a GLMM with crossed random factors to investigate factors that could be associated with students' strategy choice in their first study attempts. Three variables were included in the fixed effects, including the initial confidence on each topic, the difficulty of each topic, and the question type. Regarding the random effects, according to the research design, the initial confidence and perceived difficulty were two within-subject factors at topic level. Concerning the student level, the initial confidence, perceived difficulty and the type of the question were within-subject factors.

The results of the model can be found in Table 19. A statistically significant main effect for confidence was found for the strategy choice (b = -0.28, p < .001), indicating that the more confident the student was on the topic, the more likely they were to choose testing. The main effect for question type (b = 4.44, p < .001) also was significant, suggesting that for multiple-choice questions, students tended to choose testing; while for short-response questions, students tended to choose rereading. However, the perceived difficulty of the topic was not significantly associated with students' strategy choices.

Table 19 GLMMs results of the first study attempt (Model 1), the repeated study decision (Model 2), and the second study attempt (Model 3).

Fixed effects	***Model 1***	***Model 2***	***Model 3***
Intercept	-0.02 (0.60)	0.16 (0.40)	-0.15 (0.74)
Confidence	-0.28 ***(0.06)	-0.16***(0.05)	-0.16** (0.06)
Difficulty	0.02 (0.05)	0.00 (0.32)	-0.03 (0.07)
Question Type	4.44 ***(0.50)	0.14 (0.28)	1.84 ***(0.40)
Prior strategy choice	n.a.	1.25 **(0.41)	-2.05 ***(0.37)
Random effects: Student			
Intercept	12.34 (3.51)	3.38 (1.84)	7.74 (2.78)
Confidence	0.08 (0.29)	0.07 (0.26)	0.01 (0.10)

Difficulty	0.01 (0.11)	0.01 (0.09)	0.00 (0.07)
Question type	7.18 (2.68)	1.04 (1.02)	0.76 (0.88)
Prior strategy choice	n.a.	4.18 (2.05)	1.75 (1.32)
Random effects: Topic			
Intercept	0.49 (0.70)	0.56 (0.75)	1.60 (1.26)
Confidence	0.00 (0.70)	0.02 (0.10)	0.00 (0.03)
Difficulty	0.00 (0.70)	0.00 (0.04)	0.01 (0.12)
Prior strategy	n.a.	1.84 (1.36)	0.32 (0.57)

Note. Standard Error in parentheses in the fixed effects part; standard deviation in parentheses in the random effects part.

3.2.2 Model 2: Students' repeated study decisions

In the second model, we investigated what factors associated with students' repeated study decisions after their first study attempts. According to the research design, students' strategy use in the first study attempt, their confidence after first attempt, and perceived difficulty were within-subject factors on topic level. Concerning the student level, the confidence, perceived difficulty and the type of the question were three within-subject factors.

As shown in Table 19, the results showed that there was a statistically significant main effect of the strategy choice in the first study attempt for the decisions of repeated study (b = 1.25, $p < .01$), suggesting that when the rereading strategy was chosen in the first attempt, students were more likely to study for the second time. Conversely, when the testing was used in the first attempt, students were less likely to review the topic for the second time. In addition, a significant main effect of the confidence after the first study attempt was found for the repeat study decision (b = -0.16, $p < .001$), which meant that when students were more confident on the topic, they were less likely to repeat their study; when students were less confident, they tended to repeat their study of the topic. The perceived difficulty of the topic and the question type were not significantly associated with students' repeat study decisions.

3.2.3 Model 3: Students' strategy choice on second study attempt

In the third model, we explored how students' second strategy choice (dependent variable) was affected by their strategy choice in first attempt, confidence on the topic after first study attempt, perceived difficulty of the topic, and type of testing questions.

Results of the GLMM revealed a statistically significant main effect for first strategy choice (b = -2.05, p < .001), indicating that students tended to choose a different strategy in their second study attempt than they did in their first one. The main effect for confidence after first attempt on each topic was also significant (b = -0.16, p < .01), suggesting that the more confident the student was on the topic, the more likely they were to choose testing. Additionally, a significant main effect for question type (b = 1.84, p < .001) was found, which indicated that students tended to choose the testing strategy to learn multiple-choice questions and choose the rereading strategy for topics that were tested by short-response questions.

3.3 Students' perceptions of the use of rereading and testing

After the intervention, students were surveyed about their perceptions of testing and rereading, as well as, in each case, the circumstances under which they would use the two strategies. Most students (89 out of 152) perceived testing as a more effective strategy, while about 24% of all students (N = 37) regarded rereading as the more effective strategy, and about 17% of all students (N = 26) believed that the two strategies should be used in combination.

3.3.1 Students' perceptions of rereading

Based on students' responses, two general categories were summarised to explain why students felt that rereading was beneficial. First, rereading was a straightforward way to review knowledge and could save more time. '*Rereading puts the knowledge right in front of*

me, making it easier to review it.' (Student #60). Through rereading key points, students could refresh their memory and understanding of previously learned knowledge, thus strengthen its retention. Second, many students mentioned that compared with testing practice, rereading previously learned content could help them build up a complete knowledge structure. '*Although doing practice tests is quicker, its effects cannot last for long time and testing has a narrow focus. In comparison, I can build a comprehensive understanding of the curriculum and get high scores after rereading the textbook before the exam*' (Student #23). Given that students had learned the knowledge and had an impression in mind, rereading the content allowed them to summarise and consolidate what had been learned and get greater clarity.

3.3.2 Conditions under which students utilise rereading

According to students' responses to what factors may affect their use of rereading, their utilisation of this strategy related to four factors: 1) length of preparation time till the test; 2) knowledge mastery; 3) characteristics of the to-be-learned knowledge; 4) in what ways the to-be-learned knowledge is being tested.

Some students mentioned that they would decide whether to use rereading based on the amount of test preparation time they would have. About 18% students ($N = 27$) would use rereading more often when there was enough time to prepare for the test; while only about 5% students ($N = 9$) reported that they would use rereading when the test date drew near.

Students' responses also suggested the degree of knowledge mastery possibly had a great impact on their strategy choices. About 40% of students ($N = 60$) reported that they would make strategy decisions according to their self-evaluation of the knowledge mastery and that they would reread text of the topic if they believed that they had a poor mastery of relevant knowledge.

Besides, the type of knowledge influences students' choices, though there was a great deal of disagreement on how the knowledge characteristics affect strategy decisions. Several students (N = 12) mentioned that they would choose the rereading strategy when the learning content was wide and unstructured. There were also 13 students reporting that they would reread the knowledge when the content was relatively narrow or when the knowledge required recall.

It seems that strategy decisions were also influenced by the way students predicted how the knowledge would be tested. Thirty students (20%) reported that they would use rereading to review those topics that were frequently tested by short-response questions (i.e., definition recall or theory critique). Five students (3%) also suggested that they would take the teacher's prediction of the likelihood of a topic to be included in the test into account and that they preferred to reread the topics that have been highly anticipated by teachers.

Table 20 Themes and codes of students' perceptions on when to use rereading**.**

	Theme	Code	N
	Time	When there is enough time	27
		When it is close to the exam dates	9
	Knowledge mastery	Poor mastery of knowledge	60
	Type of the knowledge	When the learning content is wide, unstructured, or difficult to understand	12
		When the knowledge is narrow, memory-base, or easy to understand	13
	Format	The knowledge will be tested in short-response questions	30
		When the knowledge is highly likely to be tested	5

3.3.3 Students' perceptions of testing

As mentioned above, more than half of the students believed that testing was more effective than rereading, and they preferred using testing for test preparation. In general,

students recognised both the direct and indirect effect of testing on their learning, and they acknowledged the benefits of feedback. The reasons why testing strategy was considered by students as a beneficial strategy given by students could be summarised under four subthemes.

Testing can improve memorisation of knowledge

Many students ($N = 96$) recognised the direct effect of testing, that through testing, the memorisation of knowledge can be improved. Some students believed that immersing themselves with testing could strength their memorisation of knowledge: '*Constantly answering questions allows me to remember knowledge more effectively and subconsciously.*' (Student #16); '*Answering testing questions can make a deeper impression of the knowledge… and you'll naturally know the correct answer once you have done hundreds of practice questions.*' (Student #47). Some students realised that testing required a higher cognitive demand, which led to better retention. For example, '*There will be an active thinking process during testing practice'* (Student # 41, Student #124); '*Testing requires you to concentrate on the practice*' (Student #68); '*Testing requires high cognitive demands and more active brain activities, thus I will have better retention of knowledge.*' (Student #130).

Testing can diagnose knowledge mastery

Most students ($N = 83$) mentioned that testing was an effective test preparation strategy because it provides information of their knowledge mastery. '*You will know which topics you haven't mastered or understood, then you can strengthen this less understood knowledge afterwards.*' (Student #45, #69, #77, #97). Feedback and mistakes are helpful learning resources: '*If I get it wrong, I look for an explanation in the corrective feedback and compare the difference between my incorrect answer and the correct answer, so I can remember the points more firmly.*' (Student #19).

Testing can familiarise students with test format and frequently tested content

Several students (N = 41) also highlighted that testing practice could help them to get familiar with the test format and '*have a clear idea of how questions in the final course test will be constructed*' (Student #27). Thus, students may know how to apply what they have learned to answer test questions in exams. Besides, testing could be a time-saving way to prepare for tests, because '*You will know which knowledge is frequently tested, and you will be able to master the key points through testing practice rather than going through tedious texts'* (Student #86). Students also mentioned that they can improve their test-taking skills through testing practice: '*After immersing myself in frequent testing practice, I can know at a glance what the multiple-choice question is about and eliminate the impossible options to identify the most certain answer*' (Student #21); '*Testing practice allows you to find out which points can get scored. Thus, I may manage to get more scores even if I am not certain about the question in formal exams*' (Student #53).

Positive feelings about testing

Interestingly, some students (N= 14) mentioned that they experienced positive emotions from testing practice. For example, when they answered the questions correctly, they experienced a feeling of accomplishment, while if they answered it wrong, they felt motivated to learn from the feedback and put more effort to this topic.

3.3.4 Conditions under which students utilise testing

Students were also asked to describe the conditions under which they would use testing. Themes and codes of students' responses are summarised in Table 21.

About one third of students (N = 48) said that they would use testing when the time was very limited, while a relatively small number of students (N = 7) mentioned that they would use testing as a way to improve their knowledge memory when there is adequate time before the final test.

Students' knowledge mastery seemed to be an important factor, as demonstrated by the fact that about 43% of students ($N = 66$) mentioned that they would use testing to reinforce their learning of knowledge they already knew (or knew a bit). There were 12 students who said that they would use testing to learn the knowledge they did not know or to diagnose their knowledge mastery, particularly to assess knowledge they thought they had mastered.

The characteristics of the knowledge also had an impact on students' use of testing. About 12% of the students ($N = 18$) replied that they would use the testing strategy when the to-be-learnt content was wide or complicated. While about 7% of students ($N = 11$) mentioned that they would test themselves on the easy knowledge and that which did not require deep understanding.

Some students ($N = 20$) also stated that they would use testing to practice multiple-choice questions more often, as this could be a quick way to practice and diagnose their knowledge mastery. Besides, some students ($N = 8$) mentioned that they would choose testing when they needed to familiarise themselves with the upcoming summative test and practice test-taking skills, which would improve their test performance.

Table 21 Themes and codes of students' perceptions on when to use testing.

Theme	Code	N
Time	When it is close to the exam dates	48
	When there is enough time	7
Knowledge mastery	When students know the knowledge	66
	When students don't know the knowledge, don't know how to apply the knowledge, or want to diagnose their knowledge mastery	12
Type of the knowledge	When the knowledge to be learned is wide, abstract, or complicated,	18
	When the knowledge to be learned is easy, clear, or don't need deep understanding	11
Ways being tested	When the knowledge is frequently tested by multiple-choice questions	20
	When students want to familiarise themselves with the test and practice test-taking skills	8

4. General Discussion

In this research, we examined how students monitored and controlled their use of two test preparation strategies, rereading and testing, in a test preparation session before the course test. Students' perceptions of the effectiveness of strategies and their strategy use were also investigated, in order to help us better interpret students' behavioural data. This study depicts students' study behaviours in authentic test preparation contexts. Although some previous experimental studies (Badali et al., 2022; Toppino et al., 2018; Tullis et al., 2018) explored students' use of the two strategies, they were mainly conducted in laboratory settings where word pairs were used as the learning materials. These experiments cannot reflect real life test preparation situations, where materials are usually more complicated and require not only knowledge recall, but also understanding and application.

Test preparation is a unique learning occasion where individuals are required to comprehend and retain the material to the best of their abilities in a time-bound manner. The process of test preparation usually has a heightened focus and requires a greater degree of concentration and attention to detail, as the pressure to perform well on the exam can motivate students to engage more and study efficiently (Bishop, 1997; McDaniel et al., 2007). Therefore, it is important to understand how students utilise and reflect on the use of different strategies under such pressured learning conditions. Findings from this study can advance our understanding of students' metacognitive monitoring and control processes, which underlie self-regulated learning in ecologically valid contexts. Additionally, by analysing students' strategy choices and their perceptions, we could gain a better understanding of how high-stakes tests affect students' learning.

4.1 Chinese students' use of testing and rereading in test preparation contexts

In the current study, students were allowed to select either rereading or testing to study topics in educational psychology, and they could make as many study attempts as they wanted. We

recorded students' strategy choices on every study attempt and found that participants in this study tended to use testing and rereading relatively equally in their first attempts. The proportion of testing was higher than findings reported in previous studies with a similar design such as Jane et al. (2018). They investigated students' use of self-testing across sessions in a test preparation context and identified that testing accounted for a small proportion of the activities across test preparation sessions, especially at the first session (12% in study 1 and 17% in study 2) when students just studied all the materials prior to their self-study session . They held a second session two days after the first one, and the proportion of first strategy choices that were testing even decreased (7% in study 1 and 8% in study 2). We believe that the provision of the test preparation session in our study was analogous to the first and second sessions in Jane et al.'s study, as students had just finished all their curriculum and were starting to prepare for the final course exam. In our study however, students' familiarity with the knowledge may be lower than that in Jane et al.'s study, as some of the topics were learned weeks or even months earlier and as a result their retention may be reduced.

However, we observed a proportion of 48.6% the first attempts being testing, and about 70% of all testing trials were not followed by any additional learning attempts, making *using testing to study the topic for one time* the most frequent pattern. These findings are highly indicative of the preference and value Chinese students place on testing. Existing literature mainly claimed that students are inclined not to use testing because they do not consider testing as an effective learning tool or do not enjoy the process of self-testing (Carpenter et al., 2020; Geller et al., 2017). Students hardly recognise the direct effect of testing on learning and usually use testing as a monitoring tool to diagnose their learning (for a review, see Rivers, 2021). However, this notion is not supported by the current study. According to students' responses in our study, a large proportion of participants acknowledged the benefits

of test-enhanced learning, meanwhile students tended to believe that testing is more efficient than rereading.

We suggest that the value students place on a particular strategy can also be explained from a socio-cultural perspective. Existing studies on the self-regulated use of testing and rereading were all conducted in western contexts, mainly in the U.S., while none of the relevant studies were conducted in an East Asian context. Students' study habits are ingrained through years of formal education, and their study beliefs are also affected by socio-cultural factors (Schommer-Aikins & Easter, 2008; Turingan & Yang, 2009; Zhou & Wang, 2022). Taking into account the differences in educational systems between Western and East Asian countries, and the differences in social attitudes towards testing, it is not surprising that students from East Asian countries use and regard testing and rereading in different ways. The current study was conducted in China, where high-stakes testing has been used pervasively in all aspects of the society, not only for educational selection, allocation and monitoring from primary to postgraduate levels, but also for the employment in the civil service and various professional qualifications. China's educational system is well known for its examination-conscious and high-pressure nature, and most students are accustomed to the presence of frequent examinations (Kipnis, 2011; W. Li & Li, 2010). Testing has been widely considered one of the highways to academic success. There is a Chinese idiom that has been long-used to characterise the phenomenon of using testing as a way to gain good grades – '题海战术' (Ti Hai Zhan Shu, which means '*Excessive Testing*'). Besides, in recent years, a growing number of students from rural backgrounds who get into elite universities through their hard working are calling themselves 'small-town swot'[2] or 'small town test-takers'[3] (小

[2] This translation was suggested by The Economics (*Education in China Is Becoming Increasingly Unfair to the Poor*, 2021).
[3] This is the word-to-word translation.

镇做题家, Xiao Zhen Zuo Ti Jia). Though the two idioms are to some degree pejorative, they demonstrate that the effect of testing activities on achieving academic achievements is appreciated by a large population in China. The high esteem of testing in Chinese society, from the deep embeddedness of testing in social mechanisms, to the recognition of the effectiveness of testing as a learning strategy by stakeholders, explains why students in our study showed a preference for testing as a test preparation strategy in terms of both behaviours and perceptions. A possible direction for future research would be to provide more evidence on students' perceptions of different learning strategies from a comparative perspective, to contribute to our understanding of how students' study habits and beliefs are socially constructed.

Regarding students' use of rereading, it was found about 50% of students' first study attempts were rereading trials, however, this proportion decreased significantly in following study attempts. Individual participants' study attempts on each topic were counted and analysed. It was found that the first strategy choice could predict the overall study attempt students made on the topic. Students tended to revisit the topic more often when they initially reviewed it by rereading. When they initially chose to test themselves on the topic, they tended to visit the topic for fewer times. The present study identifies that the most frequent pattern adopted by Chinese students was the sole use of testing, followed by using both testing and rereading in turn. Students rarely used rereading repeatedly to learn the same topic. The findings suggest that participants tended to shift to testing in later learning attempts after getting familiar with the topic in the first study attempt. This conclusion is widely observed in previous research in terms of learning of word pairs (Badali et al., 2022; Kornell & Son, 2009) or definitions of key-concepts or texts (Janes et al., 2018; Kuhbandner & Emmerdinger, 2019). It is believed that the main reason why students abandon rereading and choose self-testing in later learning stages is because they perceive rereading as a direct

learning tool and testing as a monitoring tool (Ariel & Karpicke, 2018; Janes et al., 2018; Rivers, 2021). Our analysis of factors associated with strategy choices can provide insights on this question which we discuss below.

4.2 How do students make decisions about strategy choice?

The results of generalised linear mixed effect models demonstrated that in both first and second study attempts, students' study decisions about how to study each topic (i.e., what strategy to choose) were predicted by both their confidence on the topic and how the topic would be tested. Specifically, in our study, students tended to choose testing on topics they perceived as more well learned and choose to reread the texts of topics they perceived as poorly learned. The analysis of students' responses to the survey confirmed that students' evaluations of their own learning determined which strategy students would like to use. Most students stated that they would reread the topic when their knowledge mastery was weak and use testing to reinforce the learning of well-learned knowledge, while a small percentage of students held opposite views. This finding echoes Karpicke (2009) and Toppino et al. (2018, Experiment 1), as they also found that students' judgement of learning affects their use of testing.

Our study has reproduced many of the findings reported in the literature. However, one outcome was unexpected: although students' ratings of the difficulty of each topic negatively correlated with their ratings of confidence, the perceived difficulty was not a significant predictor of students' strategy choices. This finding seems inconsistent with the results reported by Tullis et al. (2018) and Badali et al. (2022), who stated that students regulate their use of self-testing as a function of item difficulty. Differences between the current study and prior ones in terms of the materials used and how the difficulty of materials was manipulated may be the cause of the different findings. Tullis et al. (2018) and Badali et al. (2022) used word pairs as materials in their experiments, and they were able to manipulate the difficulty

by providing students with a set of normatively easier and normatively difficult word pairs. In the current study, we were unable to identify the normative difficulty of educational psychology topics chosen from the textbook and teaching slides, instead, we asked students to rate the difficulty of each topic as objectively as they could. The finding suggests that students choose strategies based on the monitoring of their own knowledge level rather than the perceived difficulty of the knowledge. In other words, if students are confident in their mastery of a topic, even if they regard it as difficult, they may tend to choose testing to review the topic.

In terms of the impact of question type on students' strategy choice, it was found that students tended to use self-testing on topics tested by multiple-choice questions, while they were reluctant to choose testing for topics that were tested by short-response questions. This model result was further corroborated with qualitative data. Students' responses indicated that most Chinese students would use rereading to review topics that may be tested by short response questions. The retrieval effort required by different question types might explain this finding. Answering short-response questions might be regarded as more effortful, and the cognitive demands required by short-response questions are perceived as higher than those of multiple-choice questions; thus students chose to use rereading to learn the topic. In addition, test preparation time seems to be another important factor that influences students' decisions to use the particular strategy on the particular question type. Students mentioned that answering short-response questions takes more time, especially when they are not sure about the answers, while answering multiple-choice questions is more time-efficient and allows them to cover more items within a limited time frame. Given that the self-paced test preparation session we provided was time restricted and students could only study for no more than 60 minutes, they may wish to go through more materials using a time efficient strategy.

4.3 Limitations

The biggest limitation is about the generalisability of the results. This study was conducted in a teacher training college in southwest China, where some students enrolled in this school with higher *Zhongkao* (i.e., the high school entrance examination) scores because of the *Free Teacher Education* policy, while other self-funded students had relatively lower *Zhongkao* score and could not get admitted to key high schools. Therefore, there was a certain level of variation of students' ability levels in this study, which guaranteed a degree of representativeness of the population. However, this college is not a typical high school and the motivation of students in this study may be different from that of students in regular high-schools or universities. When generalising the findings of this research to the wider population of Chinese students, it is necessary to take into account the context in which this research was conducted. Future research which includes a more representative sample from regular schools or universities in academic tracks can further add to findings of this research and develop a broader understanding of test preparation effect in China's contexts as well as Chinese students' behaviours and perceptions.

5. Conclusion

When prepared for an upcoming educational psychology test, Chinese students coordinated the use of testing and rereading. When students initially reviewed the material, rereading and testing were relatively equally used. Testing was used more frequently when students revisited the materials. The strategy of rereading was rarely used repeatedly and was often accompanied by testing afterwards. Students usually used rereading on topics they perceived as better learned or that would be assessed by multiple-choice questions, while they chose to reread the texts of topics they perceived as poorly learned or that would be assessed by short-response questions. Perceived difficulty of the topic did not affect test preparation strategy

selection significantly. Chinese students can recognise both the direct (testing as an effective learning tool) and indirect effect of testing (testing a monitoring tool to diagnose learning), which might relate to wide recognition of the benefits of testing practice in Chinese educational system and the ubiquitous use of testing in Chinese society.

References

Anthenien, A. M., DeLozier, S. J., Neighbors, C., & Rhodes, M. G. (2018). College student normative misperceptions of peer study habit use. *Social Psychology of Education*, *21*(2), 303–322. https://doi.org/10.1007/s11218-017-9412-z

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*(1), 43–56. https://doi.org/10.1037/xap0000133

Badali, S., Rawson, K. A., & Dunlosky, J. (2022). Do Students Effectively Regulate Their Use of Self-Testing as a Function of Item Difficulty? *Educational Psychology Review*, *34*(3), 1651–1677. https://doi.org/10.1007/s10648-022-09665-6

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bishop, J. H. (1997). The Effect of National Standards and Curriculum-Based Exams on Achievement. *The American Economic Review*, *87*(2), 260–264.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, *64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*(6), 784–792. https://doi.org/10.1080/09658211.2016.1221974

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Carpenter, S. K., Endres, T., & Hui, L. (2020). Students' Use of Retrieval in Self-Regulated Learning: Implications for Monitoring and Regulating Effortful Learning Experiences. *Educational Psychology Review*, *32*(4), 1029–1054. https://doi.org/10.1007/s10648-020-09562-w

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition* (pp. ix, 334). Sage Publications, Inc.

Education in China is becoming increasingly unfair to the poor. (2021). Education in China Is Becoming Increasingly Unfair to the Poor | the Economist. https://www.economist.com/china/2021/05/27/education-in-china-is-becoming-increasingly-unfair-to-the-poor

Geller, J., Toftness, A., Carpenter, P., Manz, C., Coffman, C., & Lamm, M. (2017). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory*, *26*, 1–9. https://doi.org/10.1080/09658211.2017.1397175

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

Hora, M. T., & Oleson, A. K. (2017). Examining study habits in undergraduate STEM courses from a situative perspective. *International Journal of STEM Education*, *4*(1), 1. https://doi.org/10.1186/s40594-017-0055-6

Janes, J. L., Dunlosky, J., & Rawson, K. A. (2018). How Do Students Use Self-Testing Across Multiple Study Sessions When Preparing for a High-Stakes Exam? *Journal of Applied Research in Memory and Cognition*, *7*(2), 230–240. https://doi.org/10.1016/j.jarmac.2017.11.003

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486. https://doi.org/10.1037/a0017341

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009

Kipnis, A. B. (2011). Governing Educational Desire: Culture, Politics, and Schooling in China. In *Governing Educational Desire*. University of Chicago Press. https://doi.org/10.7208/9780226437569

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived

learning relate to study strategy choice. *Cognitive Psychology*, *115*. https://doi.org/10.1016/j.cogpsych.2019.101237

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. https://doi.org/10.1080/09658210902832915

Kuhbandner, C., & Emmerdinger, K. J. (2019). Do students really prefer repeated rereading over testing when studying textbooks? A reexamination. *Memory*, *27*(7), 952–961. https://doi.org/10.1080/09658211.2019.1610177

Li, W., & Li, Y. (2010). An Analysis on Social and Cultural Background of the Resistance for China's Education Reform and Academic Pressure. *International Education Studies*, *3*(3), 211–215.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257–271. https://doi.org/10.1080/09658211.2014.1001992

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

Rivers, M. L. (2021). Metacognition About Practice Testing: A Review of Learners' Beliefs, Monitoring, and Control of Test-Enhanced Learning. *Educational Psychology Review*, *33*(3), 823–862. https://doi.org/10.1007/s10648-020-09578-2

Schommer-Aikins, M., & Easter, M. (2008). Epistemological beliefs' contributions to study strategies of Asian Americans and European Americans. *Journal of Educational Psychology*, *100*, 920–929. https://doi.org/10.1037/0022-0663.100.4.920

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval

practice. *Memory & Cognition*, *46*(7), 1164–1177. https://doi.org/10.3758/s13421-018-0828-2

Tullis, J. G., Fiechter, J. L., & Benjamin, A. S. (2018). The efficacy of learners' testing choices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 540–552. https://doi.org/10.1037/xlm0000473

Turingan, J. P., & Yang, Y.-C. (2009). A Cross-Cultural Comparison of Self-Regulated Learning Skills between Korean and Filipino College Students. *Asian Social Science*, *5*(12), p3. https://doi.org/10.5539/ass.v5n12p3

Winne, P. H., & Hadwin, A. F. (1998). Studying as Self-Regulated Learning. In *Metacognition in Educational Theory and Practice*. Routledge.

Yan, V. X., Thai, K.-P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition*, *3*(3), 140–152. https://doi.org/10.1016/j.jarmac.2014.04.003

Zhou, J., & Wang, X. (2022). The Influence of Chinese College Students' Meta-Cognitive Strategy and Motivational Beliefs on their Study Habits. *The Asia-Pacific Education Researcher*, *31*(2), 93–103. https://doi.org/10.1007/s40299-020-00541-x

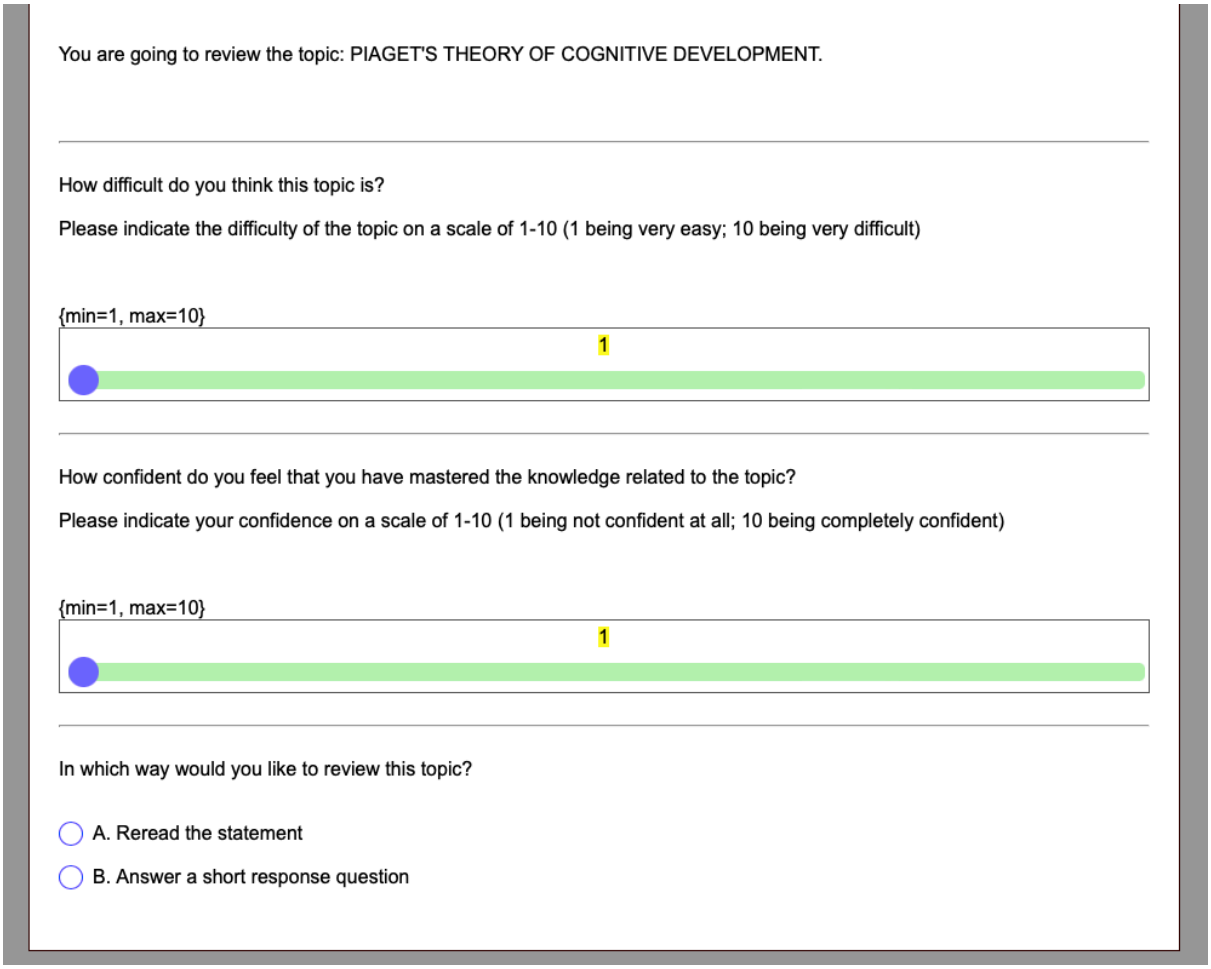
Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

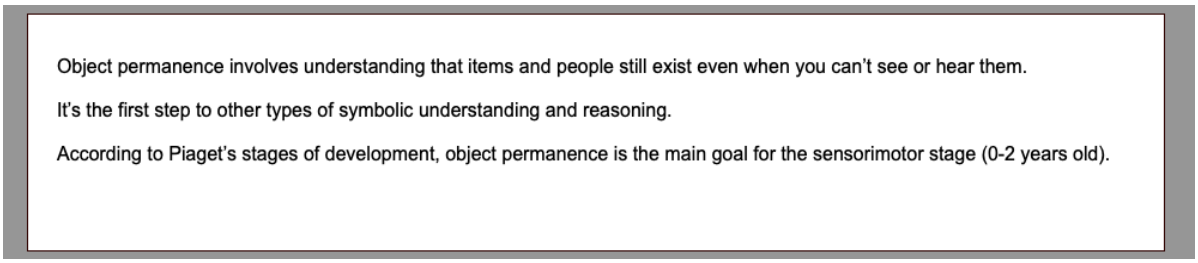
Appendix A. Examples of practice questions and statements used in the intervention

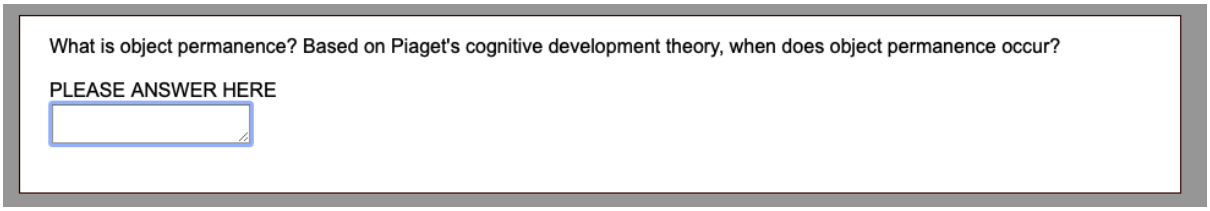
Concepts	Testing	Rereading
Zone of proximal development	Question: Please explain the concept of ZPD. Correct Feedback: ZPD is the difference between what a child can do alone and what child can do with the help of a teacher, includes a child's emerging skills.	ZPD: the difference between what a child can do alone and what child can do with the help of a teacher, includes a child's emerging skills.
Piaget's cognitive development theory	**Question:** A child named Ming can deduce the relationship between A, B and C based on the principle that A is bigger than B, B is bigger than C, then A is bigger than C. For example, 'Zhang is taller than Li, Li is taller than Wang, so Zhang is the highest.' According to Piaget's cognitive development theory, Ming is at _____ stage. A. Sensorimotor stage B. Preoperational stage C. Concrete operational stage D. Formal operational stage	Ming can deduce the relationship between A, B and C based on the principle that A is bigger than B, B is bigger than C, then A is bigger than C. For example, 'Xiao Zhang is taller than Xiao Li, Xiao Li is taller than Xiao Wang, so Xiao Zhang is the highest.' According to Piaget's cognitive development theory, Ming is at Formal operational stage.

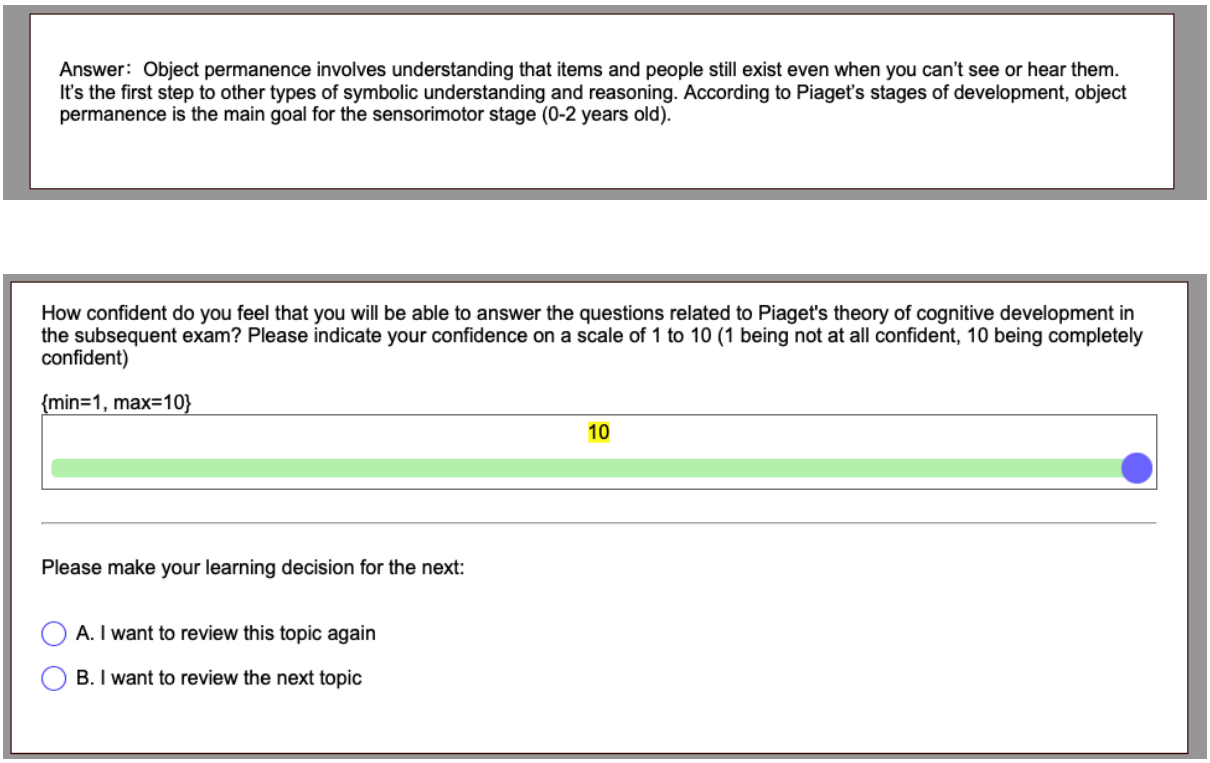
	Feedback: The correct answer is D.	
Reinforcement	**Question:** What is the difference between cognitive strategies and metacognitive strategies? Please give a brief description with examples. Feedback: Cognitive strategies are specific methods/techniques of information processing, which help effectively solve problems or complete tasks. Cognitive strategies include repetition, organizing and summarizing strategies. For example, copying, taking notes, making outlines, making associations, etc. Metacognitive strategy is the cognitive strategy students used on their cognitive process, including the monitor and control strategy of their own cognitive process, which helps students effectively arrange and regulate the learning process. Metacognitive strategies include planning strategies, monitoring strategies and adjustment strategies. For example, setting learning goals, paying attention, taking self-tests, etc.	Cognitive strategies are specific methods/techniques of information processing, which help effectively solve problems or complete tasks. Cognitive strategies include repetition, organizing and summarizing strategies. For example, copying, taking notes, making outlines, making associations, etc. Metacognitive strategy is the cognitive strategy students used on their cognitive process, including the monitor and control strategy of their own cognitive process, which helps students effectively arrange and regulate the learning process. Metacognitive strategies include planning strategies, monitoring strategies and adjustment strategies. For example, setting learning goals, paying attention, taking self-tests, etc.

Appendix B. Screenshots of intervention tasks









Note. The screenshots illustrate the rating, decisions, and learning tasks included in each study attempt. The top panel illustrates the monitoring and control tasks before learning the topic; the middle panels illustrate students learning the topic via either rereading or testing; the bottom panel illustrates the monitoring and control task after one study attempt.

Appendix C. Model fit value

Table C1. Model fit value of the first model

	npar	AIC	BIC	LogLik	deviance	Chisq	df	P
Model 1_{null}	3	2104.3	2121.7	-1049.15	2098.3			
Model 1	26	2035.6	2308.0	-970.78	1941.6	156.75	44	<.001

Note. Model 1_{null} is the model with only random intercept; Model 1 is the model with random intercept and random slope.

Table C2. Model fit value of the second model

	npar	AIC	BIC	LogLik	deviance	Chisq	df	P
Model 2_{null}	3	2872.0	2889.4	-1433.0	2866.0			
Model 2.2	26	2644.6	2795.3	-1296.3	2592.6	273.36	23	<.001

Note. Model 2_{null} is the model with only random intercept; Model 2 is the model with random intercept and random slope.

Table C3. Model fit value of the third model

	npar	AIC	BIC	LogLik	deviance	Chisq	df	P
Model 3_{null}	**3**	969.76	984.62	-481.88	963.76			
Model 3	26	940.44	1069.19	-444.22	888.44	75.32	23	<.001

Note. Model 3_{null} is the model with only random intercept; Model 3 is the model with random intercept and random slope.

Author Contributions

Zhanxin Hao: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing-Original Draft.

Jo-Anne Baird: Conceptualization, Writing-Reviewing & Editing, Supervision

Yasmine El Masri: Conceptualization, Writing-Reviewing & Editing, Supervision

Part III
Synthesis

Chapter 7. Overview of findings and additional analysis 238

Chapter 8. General Discussion 280

Chapter 7. Overview of findings and additional analysis

This chapter starts with a brief summary of findings from the three research papers in Part II. Then, some new findings are introduced based on additional analyses of data that are not reported in the papers. Finally, this chapter provides an integrated overview of the complex findings in this research.

7.1 Summary of findings of the three papers

The three papers in Part II of the thesis provided rich and valuable findings on the average effect of test preparation interventions published in the literature (Paper 1), effects of the use of testing and restudy in both high-stakes and low-stakes settings and the add-on effects of metacognitive monitoring and control to testing and restudy practices (Paper 2), and Chinese students' study behaviours during self-led test preparation (Paper 2 and Paper 3), as well as factors associated with students' strategy use and students' perceptions of testing and restudy practices (Paper 3).

The three research papers revealed some important findings. The first paper is a meta-analysis which quantifies the average effect of test preparation on large-scale educational tests based on previous (quasi-) experimental studies. In general, a significant and positive test preparation effect on large-scale educational tests was found ($g = 0.38$, $p < .001$). Furthermore, moderating variables that were more likely to affect the magnitude of effect size were identified (see Figure 14 for a visual summary). The results suggested that the breadth of test preparation intervention, the use of workbooks, test preparation management skills and teaching test-taking skills were significant moderators of test preparation effect. Besides, it seems that test preparation interventions on EFL tests had greater effects than interventions on verbal tests, and students who had previous experience of test-taking and test preparation benefited more from interventions than those who did not have any previous exposure.

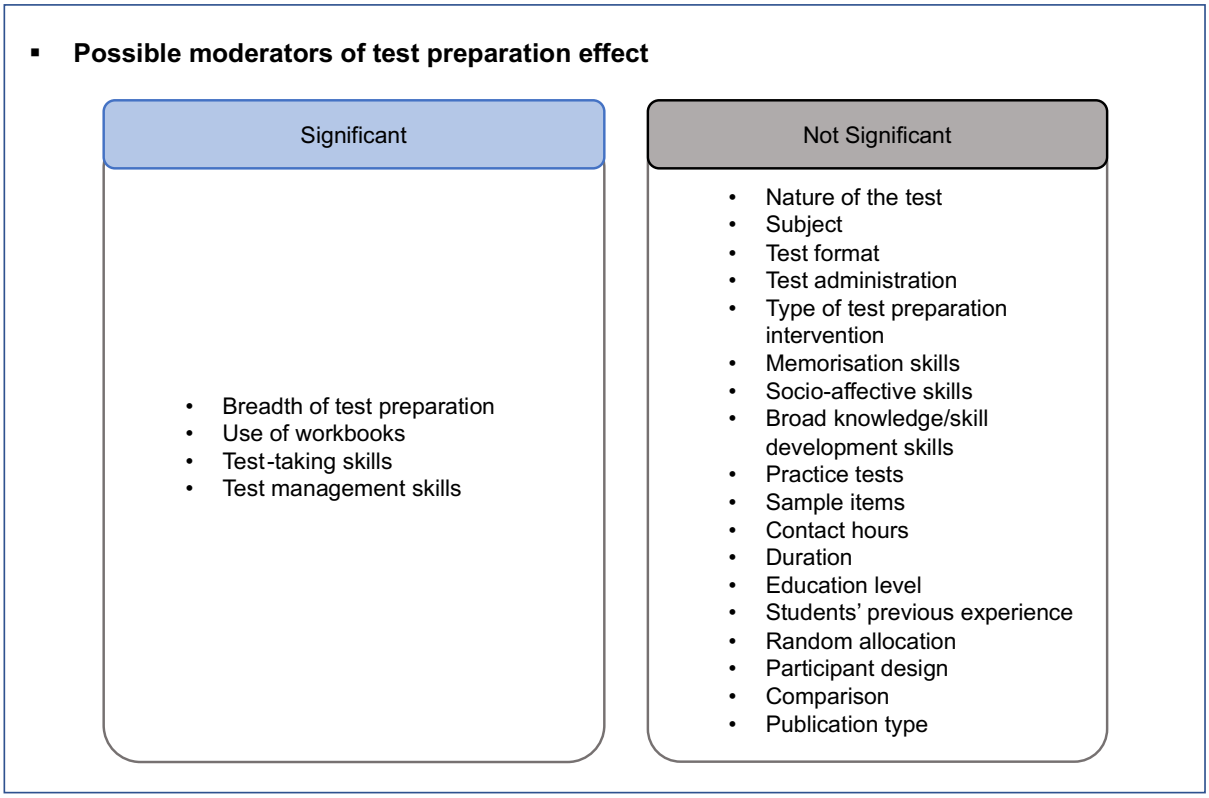


Figure 14 Significant and non-significant moderators of test preparation effects in the meta-analysis in Paper 1.

As shown in Figure 14, the meta-analysis in Paper 1 did not provide enough support for the superior effects of taking practice tests and sample items during test preparation than other practices (e.g., lectures), and this contradicts with findings reported in a large body of memory and retention literature (e.g., Agarwal et al., 2021; Schwieren et al., 2017; Rowland, 2014). Based on this finding, Paper 2 investigates whether a testing effect exists in test preparation contexts where authentic assessment tasks usually require a certain level of knowledge transfer from study materials. Three (quasi-) experimental studies were introduced in Paper 2. To aid the interpretation of complex findings across the thesis, some block diagrams are used (see Figure 15 and Figure 16) to indicate significant differences between conditions. This strategy is used here and in the additional analyses presented later.

Figure 15 presents the significant differences in post-test performance between groups across Experiments 1 to 3 based on ANCOVA models, with post-test score as the dependent

variable, pre-test score as the covariate, and the intervention groups in each experiment as independent variables. The experimental interventions resulted in significantly better post-test performance only in Experiment 1 and Experiment 3 (i.e., high-stakes contexts). In Experiment 1, only the *Testing with feedback* and *Restudy* groups produced significantly better post-test scores than the control group. In Experiment 3, all experimental conditions resulted in a higher post-test performance than the control group. In Experiment 2 conducted after the final course test, none of the intervention groups led to significantly higher post-test performance than the control group.

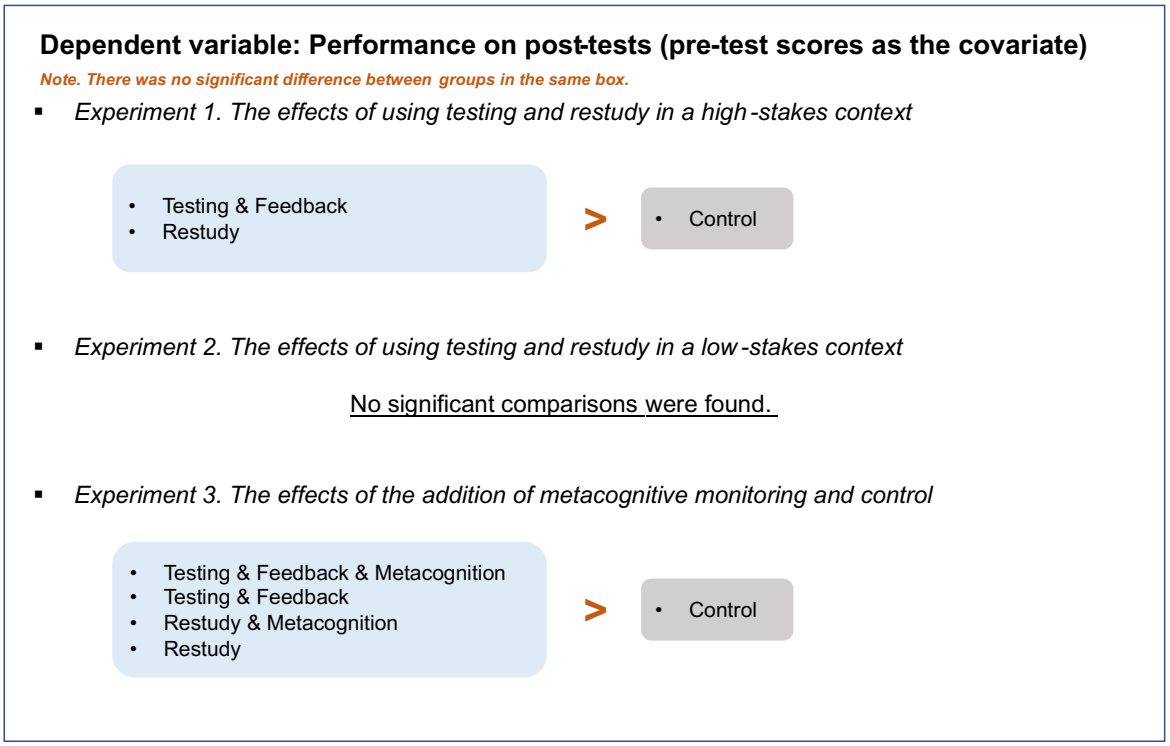


Figure 15 Significant differences from ANCOVAs on **post-test performance** between test preparation conditions across Experiments 1 to 3.

As for students' performance on course tests, significant results from ANCOVAs are presented in Figure 16, with final course test score as the dependent variable, pre-test score as the covariate, and the intervention groups in each experiment as the independent variables. In Experiment 1, only the *Testing with feedback* group had significantly better post-test scores

than the control group. In Experiment 3, the *Testing with feedback*, *Testing with feedback plus metacognition*, and *Restudy plus metacognition* produced significantly higher course test performance than the control group. The analysis was not applicable in Experiment 2, as it was conducted after the final course test and no delayed test was used after the post-test.

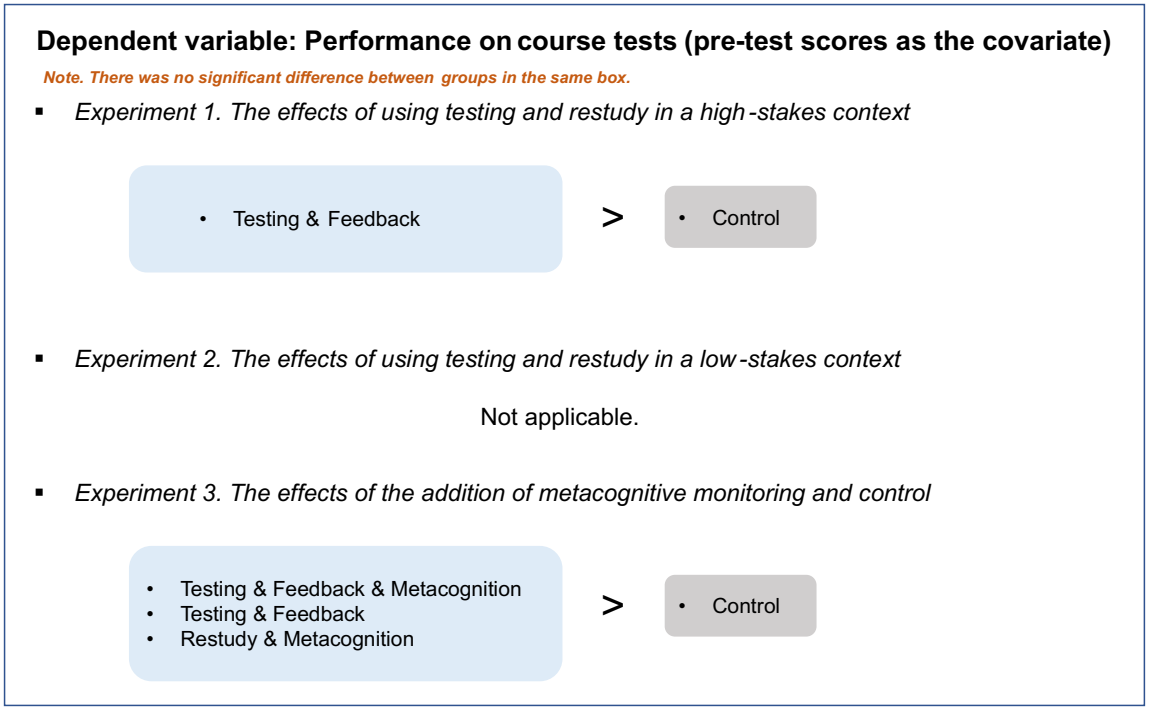


Figure 16 Significant differences from ANCOVAs on **course-test performance**.

In Paper 3, students were allowed to choose either testing or restudy to review educational psychology topics. They were asked to make judgements on topic difficulty and their confidence in study for each topic. After the intervention, students answered a series of open-ended questions in a survey to share their perceptions of the two strategies. Results of Paper 3 suggested that the Chinese students in the study most frequently chose one-time testing to study a topic, followed by the combined use of testing and restudy, while repeatedly restudying a topic was the least frequent practice. Students' confidence in a topic and how the topic was tested were two significant predictors of students' strategy choices. Specifically, students tended to choose testing when they felt more confident and when a

multiple-choice question was provided if students chose to test themselves on the topic, conversely, students tended to choose to restudy when they felt less confident in the topic and when a short-response question was provided if students chose to test themselves on the topic. Qualitative data also confirmed students' preferences for testing practices. The analysis suggested that students recognised both the direct benefits of testing practices on learning and the indirect effects of testing, i.e., as a monitoring tool. Meanwhile, students tended to believe that restudy was effective only in combination with testing practices.

To note, some data collected in the four (quasi-) experiments were not covered by the three papers. Due to the length restrictions on journal papers, each paper was written using a central narrative thread to connect the most relevant results. Thus, some data that were deemed to be less relevant to the main narrative of the papers were not reported. For instance, Paper 3 focused on students' choices of the two strategies (testing and restudy) during their self-study, highlighting the processes of how strategies were used and why students made such study decisions. The effect of students' regulation of the two strategies was out of the scope of the paper and thus was not presented. However, these data could be valuable additions for portraying a more complete picture of test preparation and could contribute to a better interpretation of findings. Therefore, some new analyses are introduced in subsequent sections in this chapter to present these data that were not reported in papers. Furthermore, since the four (quasi-) experimental studies are interrelated, some cross-experimental analyses are conducted to enhance the coherence of findings across these studies.

To be specific, the first additional analysis presented in this chapter follows on from Paper 3. It investigates the effect of students' combined use of testing and restudy in Experiment 4, compared with the effects of the singular use of either testing or restudy strategies in Experiment 3. Given that students in the two experiments were recruited from the same cohort of students taught by the same course teacher, and materials used in the two experiments were exactly same, this cross-experimental comparison allows an overall

understanding of the effectiveness of using each strategy specifically and using the two in combination.

The second additional analysis synthesises the effects of test preparation interventions reported across all of the four (quasi-) experiments by a continuously cumulating meta-analysis (CCMA). An overall effect of test preparation interventions in the current research programme is reported and possible moderators of the effect are identified.

Some previous literature suggested that students' prior attainment might affect the effect of test preparation (e.g., Carpenter et al., 2016; Sesnowitz et al., 1982), so the relationship between students' prior attainment level and their score gains after the experimental interventions in this research is analysed to extend the understanding of whether and how individual differences affect the effect of test preparation.

In addition, whether specific study behaviours and students' monitoring accuracy are associated with varying degrees of benefit for students with different prior attainment level has been investigated, based on analyses compared students' study behaviours (i.e., overall study attempts and attempts of using specific strategies) during the online test preparation session and changes in monitoring accuracy after the interventions among different prior attainment levels.

In a nutshell, these four sets of additional analyses improve the completeness of the overall analysis and can contribute to an overall understanding of the test preparation effects revealed by the four (quasi-) experiments in the research programme, as well as point out new aspects of understanding test preparation based on the lens of individual difference.

7.2 The effect of students' combined use of strategies in Experiment 4

As mentioned earlier, the effect of test preparation activities on the immediate post-test and the course test in Experiment 4 was out of the scope of Paper 3, as the paper mainly aimed to contribute to the understanding of Chinese students' uses and perceptions of the two

strategies (restudy and testing) in a real test preparation context. However, it is also valuable to explore the effect of students' combined use of the two strategies, especially to compare whether the effect exceed the use of a single strategy restrictedly. Therefore, a series of one-way ANCOVA models were performed to compare students' performance in the post-test and the course test with students in Experiment 3, as they were recruited from the same student cohort and used identical materials. Students' test performance in Experiments 3 and 4 are presented in Table 22, and significant group differences from ANCOVA models are summarised in Figure 17.

Table 22 Test performance on pre- and post-tests and course test in Experiment 3 and 4.

Experiment	Conditions	N	Pre-test		Post-test		Course test	
			Mean	SD	Mean	SD	Mean	SD
Experiment 3	**Testing with feedback**	39	33.77	7.50	42.85	10.63	65.21	10.35
	Testing with feedback plus metacognition	38	35.13	9.70	44.71	12.19	62.95	13.02
	Restudy	39	33.49	10.11	42.00	11.50	60.21	15.21
	Restudy plus metacognition	37	33.78	9.22	43.76	8.66	61.49	12.91
Experiment 4	**Combined use of testing and restudy**	152	37.57	8.96	45.69	11.78	60.63	13.75
Control	**Control**	104	36.21	9.517	36.28	10.19	57.10	13.73

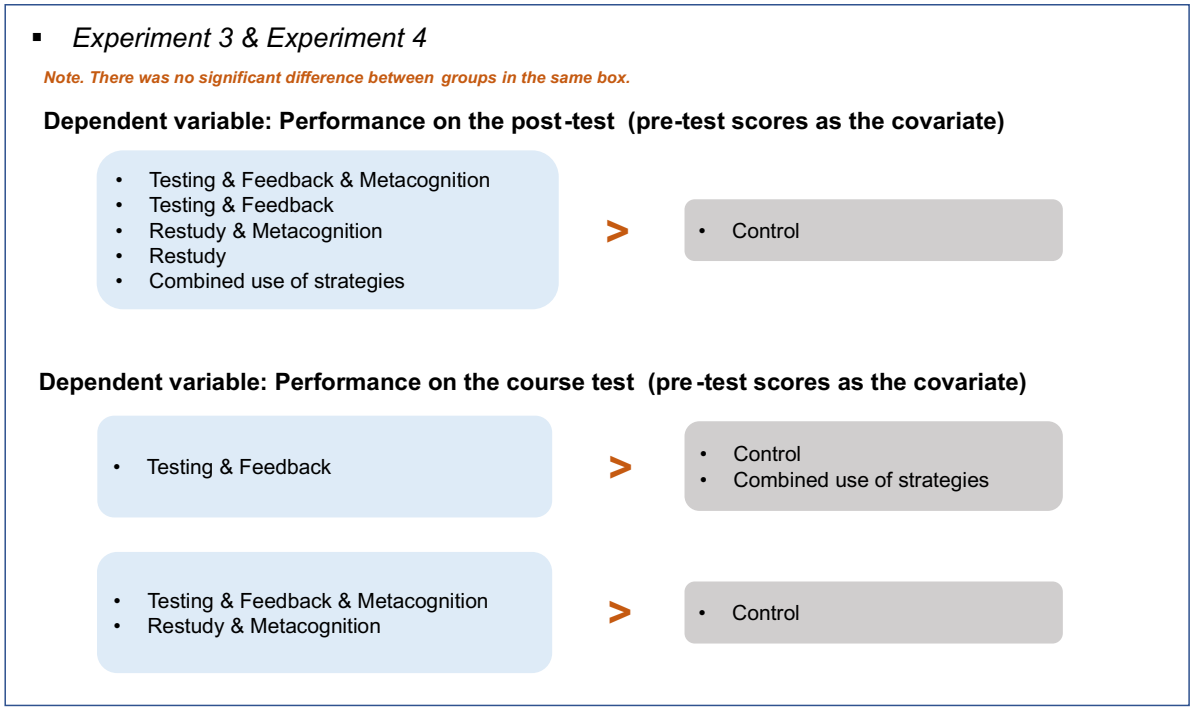


Figure 17 Significant differences from ANCOVAs on **the post-test** and **the course-test performance** across test preparation conditions in Experiment 3 and 4.

Pre-test scores

To note, though students were recruited from the same cohort, they were not randomly allocated to Experiment 3, Experiment 4 and the control condition. To examine whether the three groups of students had different levels of prior knowledge, their performance on the pre-test was compared first. The one-way ANOVA model using pre-test score as the dependent variable and the three groups (three classes in Experiment 3, three classes in Experiment 4, and two classes in the control) as the independent variable revealed that there was a significant difference in students' pre-test performance across the three groups of students, F (2, 406) = 5.75, p < .001, η_p^2 = .028. The post-hoc analysis showed that students in Experiment 4 had an averagely higher pre-test score than those in Experiment 3 (*Mean difference* = 3.53, *SE* = 1.05, p < .01). There was no other significant difference between students in Experiment 3 and those in the business-as-usual control, nor between students in Experiment 4 and in the control (ps > .05).

Due to the existence of difference in pre-test performance, pre-test scores were controlled as a covariate when analysed post-test and final course test performance. As Experiment 3 and 4 did not have same test preparation conditions, test performances of different conditions in the two experiments and the control were compared directly, i.e., the independent variable used in the following ANCOVA models on post-test and course test performance was test preparation conditions across Experiment 3, 4 and the control. In this way, precise comparisons between the effect of each specific strategy use or strategy combination could be provided.

Post-test scores

In terms of students' performance on the post-test, the one-way ANCOVA model revealed that when the pre-test score was controlled as a covariate, there was a significant difference in post-test performance across the six test preparation conditions, F (5, 402) = 14.28, p < .001, η_p^2= .151. The post-hoc analysis revealed that all the five experimental conditions in both Experiment 3 and Experiment 4 significantly outperformed the control condition (ps < .001). There was no significant difference across the five experimental conditions (ps > .05). Significant group differences in the post-test among the six conditions can be seen in Figure 17.

Course test scores

As shown in Figure 17, regarding students' performance on the final course test, the one-way ANCOVA model also identified significant group differences among the six conditions, after controlling the pre-test scores as the covariate, F (5, 402) = 6.67, p < .001, η_p^2 = .076. The post-hoc analysis reported that students in the *Combined use of strategies* group in Experiment 4 did not significantly outperform their peers in the business-as-usual control condition (p > .05). Students in the *Testing and feedback* group in Experiment 3 significantly outperformed students in Experiment 4 (*Mean difference* = 7.95, *SE* = 1.96, p <.001) and

those in the control (*Mean difference* = 10.00, *SE* = 2.04, $p < .001$). As mentioned in Paper 2, students in the other two experimental groups, i.e., *Testing with feedback plus metacognition* and *Restudy plus metacognition*, also achieved significantly better final course scores than the control ($ps < .05$, see Paper 2 for details).

Summary of the effect of Experiment 4

To summarise, findings indicated that the *Combined use of strategies* in Experiment 4 was an effective practice on students' post-test performance, and its effect was similar to the effects of *Testing with feedback* and *Restudy* practices (both with and without metacognition) in Experiment 3, which employed the same materials as Experiment 4. However, on the course test conducted one week later, students in the *Combined use of strategies* did not outperform those in the control condition. Furthermore, when compared with conditions in Experiment 3, the *Combined use of strategies* even significantly underperformed the *Testing with feedback condition* in Experiment 3. The two test preparation plus metacognition conditions (i.e., *Testing with feedback plus metacognition* and *Restudy plus metacognition*) also had significantly better course test performance than the control condition, while the *Combined use of strategies* condition did not. These findings suggested that allowing students to choose and interchange between different strategies may weaken the long-term effect on test performance, as compared to restricting students to use a single strategy. One possible reason might be that making decisions on strategy choices and switching between different strategies can be cognitively demanding and might cause extraneous cognitive load, thus negatively affect study outcomes. This explanation will be further discussed in Chapter 8.

7.3 Test preparation effect across experiments

To provide an overall estimate of the magnitude of test preparation intervention on small-scale educational tests, a continuously cumulating meta-analysis (CCMA) was further conducted as suggested by Braver et al. (2014). The CCMA synthesised data from the four

experiments and calculated an overall effect of test preparation interventions on students' performance in this research programme.

A significant mean effect of test preparation interventions was observed, $g = 0.41$, $se = 0.19$, $p < .05$. The forest plot in Figure 18 displays the effect size of each test preparation condition in every experiment.

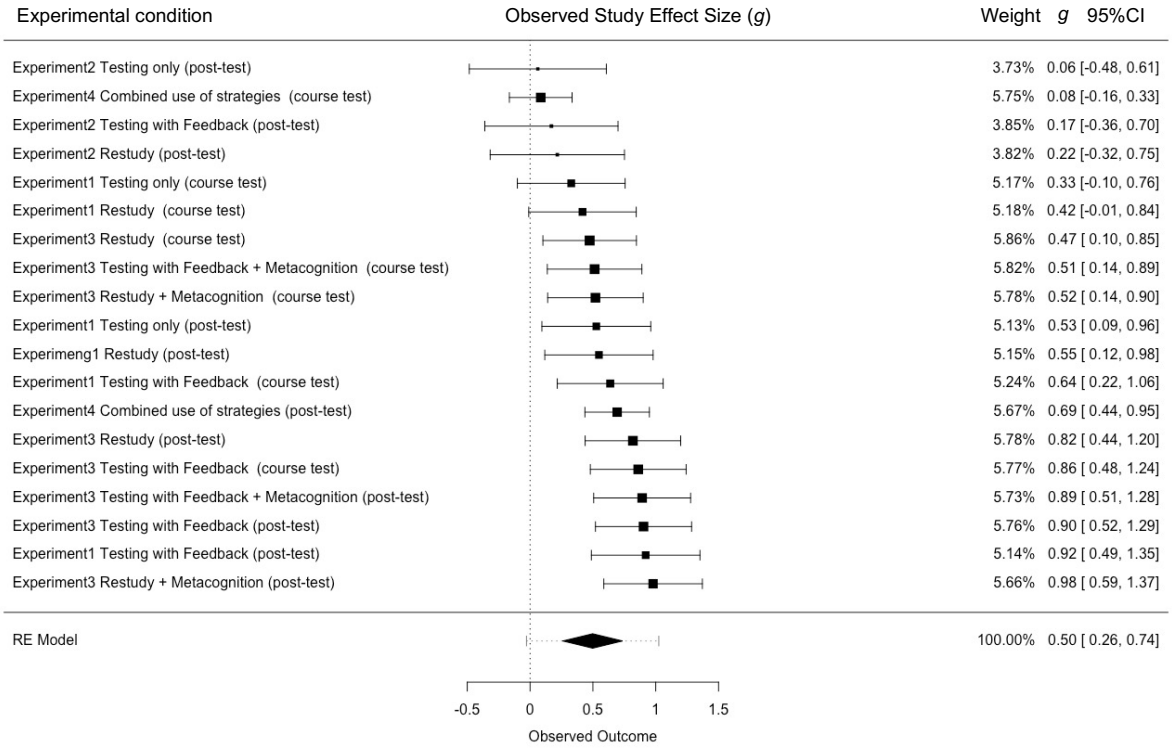


Figure 18 Forest plot of effect sizes of all test preparation interventions in the four (quasi-) experiments.

To reiterate, all four experiments were conducted in the same college in China and therefore shared some commonalities. In addition, how the test preparation effect related to different factors manipulated by each experiment has already been made clear in Paper 2. Thus, moderator analysis cannot be expected to identify new moderators of the test preparation effect based on this cumulative analysis of the four experiments. As such, a series of moderator analyses was carried out, with the aim of cross-checking the results of

Paper 2 based on the between-experiment analysis. As shown in Table 23, only the interval between the test preparation and the criterion test was found to be a significant moderator. The subgroup analysis showed that the testing preparation effect was only significant when the test was administered immediately (i.e., the post-test) (g = 0.60, se = 0.16, p <.01). According to the results of the meta-regression, test preparation interventions had a significantly smaller effect when the test was administered one week after the intervention than when the test was administered immediately, b =-0.40, se = 0.11, p <.01. In addition, test preparation strategy was not a significant moderator, but the subgroup analysis revealed some differences regarding the level of significance of different strategies. It showed that *testing with feedback* and *restudy* were two significantly effective test preparation strategies, while *testing only* and the *combined use of strategies* did not have significant effects compared to the control. These results indicated that feedback is the key to achieve a significant testing effect in authentic test preparation conditions, and the effect might be weakened by leaving students to choose their own strategies. However, caution is required, as the numbers of included cases from the two conditions were small, thus findings are not conclusive and further studies are needed. Other results of the moderator analysis (e.g., the test preparation effect was significant in high-stakes settings rather than the low-stakes setting) were consistent with findings reported in Paper 2.

Table 23 Results of moderator analysis of the CCMA.

Moderator variables	**Significance of moderator**				**Pooled Effect**						
	Q	***p* of Q**	**F**	***p* of F**	**N**	**k**	**g**	**se**	**95%CI-L**	**95%CI-U**	***p***
Test preparation strategy	**33.50**	**<.01**	**0.47**	**.707**							
Testing only					2	4	0.35	0.27	-0.23	0.92	.218

Testing with Feedback **					3	8	0.58	0.19	0.18	0.98	.007
Restudy *					3	7	0.47	0.19	0.06	0.88	.027
Combined use of strategies					1	2	0.39	0.31	-0.28	1.05	.233
Metacognitive tasks included	**40.79**	**<.001**	**0.01**	**.939**							
Yes**					2	6	0.48	0.16	0.15	0.82	.008
No**					3	13	0.50	0.14	0.20	0.79	.003
Stakes of the test	**33.66**	**<.01**	**3.12**	**.095**							
High***					3	16	0.59	0.10	0.38	0.80	<.001
Low					1	3	0.15	0.23	-0.33	0.63	.515
Interval **	**32.32**	**<.01**	**7.18**	**<.01**							
Immediate**					4	11	0.60	0.16	0.26	0.95	.002
2 days					1	3	0.41	0.22	-0.06	0.88	.080
1 week					2	5	0.21	0.18	-0.17	0.59	.259

7.4 Prior attainment levels and score gains

Previous studies suggested that students' prior attainment levels might relate to the extent to which they can benefit from test preparation interventions (Carpenter et al., 2016; McGaghie et al., 2004). Here I explore whether students' score gains are associated with their prior attainment levels in the four (quasi-) experiments conducted for this thesis. In the following analyses, students' prior achievement levels were operationalised both absolutely and relatively in terms of how their pre-test performance was compared. The first set of analyses below operationalised prior attainment in an absolute sense, with students' pre-test performance categorised across experiments. The levels of performance on the pre-test were interpreted in the same way no matter when and where it was taken. In this case, students in some experiments had higher attainment in general due to the timing of the experiment or due to the non-random allocation. The second set of analyses operationalised attainment levels relative to the group of students included in each experiment separately, who were at the

same point in terms of experience of the course and timing with respect to the final course test.

Students' score gains between the pre- and post-tests and between the pre-test and the course tests were used as dependent variables in the analyses. Score gains between the pre-test and the post-test (i.e., the post-test score minus the pre-test score), as well as score gains between the pre-test and course test (i.e., the course test score minus the pre-test score) were compared to illustrate the difference between different prior attainment levels across different test preparation practices in the four (quasi-) experiments.

7.4.1 Absolute prior attainment levels and score gains

Experiments 1 and 2 recruited participants from the same cohort of students and used the same set of test papers, and Experiments 3 and 4 recruited participants from another cohort and used another set of test papers. Therefore, students in Experiments 1 and 2 were grouped together and partitioned into three attainment levels based on whether their pre-test scores fell within the upper, middle, or lower one third. Similarly, students in Experiments 3 and 4 were grouped together and categorised into three attainment levels based on the ranking of their pre-test scores.

Score gains of students with different *absolute* attainment levels in all conditions of the four (quasi-) experiments are presented in Table 24. Note that there were few low prior attainment performers in Experiment 2 due to the absolute way in which their performances were classified (across experiments). A series of ANOVA models were performed to compare the group difference in score gains across different prior attainment levels and different test preparation conditions. Significant group differences are summarised in block diagrams in Figures 19 and 20 shown below.

Table 24 Score gains among absolute prior attainment levels in the four (quasi-) experiments.

Experiment	Test preparation conditions	Absolute prior attainment level	Score gains between the post-test and the pre-test	Score gains between the course test and the pre-test
Experiment 1	Testing & Feedback	Low (n=23)	20.48 (13.18)	28.87 (9.21)
		Middle (n=14)	3.79 (10.40)	22.07 (8.25)
		High (n=9)	-5.33 (11.17)	9.89 (9.91)
	Testing only	Low (n=15)	15.60 (12.54)	26.73 (8.21)
		Middle (n=18)	4.39 (12.07)	19.67 (6.38)
		High (n=7)	-9.00 (16.24)	7.14 (7.31)
	Restudy	Low (n=17)	11.94 (12.66)	26.53 (7.49)
		Middle (n=15)	5.67 (13.23)	19.93 (6.79)
		High (n=9)	-2.00 (9.54)	12.89 (8.18)
	Control	Low (n=17)	3.53 (6.76)	21.94 (8.45)
		Middle (n=20)	-2.15 (7.29)	16.65 (9.05)
		High (n=8)	-5.13 (8.74)	8.25 (10.83)
Experiment 2	Testing & Feedback	Low (n=6)	7.83 (11.30)	
		Middle (n=10)	9.10 (12.28)	
		High (n=16)	2.81 (10.12)	
	Testing only	Low (n=4)	3.00 (10.03)	
		Middle (n=9)	6.89 (12.55)	
		High (n=15)	2.40 (8.24)	
	Restudy	Low (n=6)	5.83 (13.26)	
		Middle (n=4)	7.50 (12.40)	
		High (n=21)	7.10 (12.40)	
	Control	Low (n=7)	11.14 (15.78)	
		Middle (n=6)	1.83 (13.17)	
		High (n=11)	-1.91 (5.17)	
Experiment 3	Testing & Feedback	Low (n=13)	9.15 (8.61)	31.92 (10.52)
		Middle (n=19)	10.79 (10.54)	31.32 (9.14)
		High (n=7)	4.29 (4.57)	30.86 (3.81)
	Testing & Feedback plus Metacognition	Low (n=11)	13.36 (8.18)	28.55 (11.41)
		Middle (n=17)	7.53 (10.47)	27.82 (9.69)
		High (n=10)	8.90 (8.83)	27.00 (7.50)
	Restudy	Low (n=15)	10.13 (6.51)	29.13 (14.96)

		Middle (n=14)	7.36 (8.72)	23.21 (13.76)
		High (n=10)	7.70 (9.97)	28.00 (10.04)
	Restudy plus Metacognition	Low (n=14)	13.71 (5.74)	28.50 (8.95)
		Middle (n=14)	10.21 (8.62)	27.07 (9.01)
		High (n=9)	3.78 (6.69)	27.44 (9.66)
Experiment 4	Combined use of testing and restudy	Low (n=41)	10.66 (9.61)	24.07 (11.93)
		Middle (n=50)	8.54 (11.86)	22.90 (10.67)
		High (n=61)	6.07 (8.55)	22.51 (11.83)
Control	Control	Low (n=42)	3.50 (7.13)	21.33 (9.46)
		Middle (n=23)	-0.52 (7.35)	21.48 (13.81)
		High (n=39)	-3.28 (9.88)	20.87 (10.35)

Experiments 1 and 2

For the first cohort of students (students enrolled in the educational psychology course during 2019-2020), random allocation to Experiment 1 and 2 was used. Within each experiment, they were randomly allocated to one of the four groups (three experimental groups and one control group, as can be seen in the Table 24 above). However, the two experiments were conducted at different times. Experiment 1 was conducted earlier, two days before the final course test, while Experiment 2 was conducted one week after the course test. The timing of the study resulted in differences in students' pre-test scores between the two experiments. When students in Experiment 2 participated in the experiment, compared with their peers in Experiment 1, they generally had greater familiarity with the course materials due to self-preparation for the course test. Therefore, variance caused by the timing of the experiment in relation to the final course test must be taken into consideration.

Three-way ANOVA models were conducted to compare group differences on score gains across three independent variables: absolute prior attainment levels, test preparation conditions, and timing of the experiment. To note, the analysis focused on the relationship between students' prior attainments and score gains, thus, only the main effect of prior attainment level and its interactions with other factors were reported. The relationship

between the other factors and score gains were addressed in Paper 2. Based on the results of ANOVA models, significant presented in Figure 19 and outlined below.

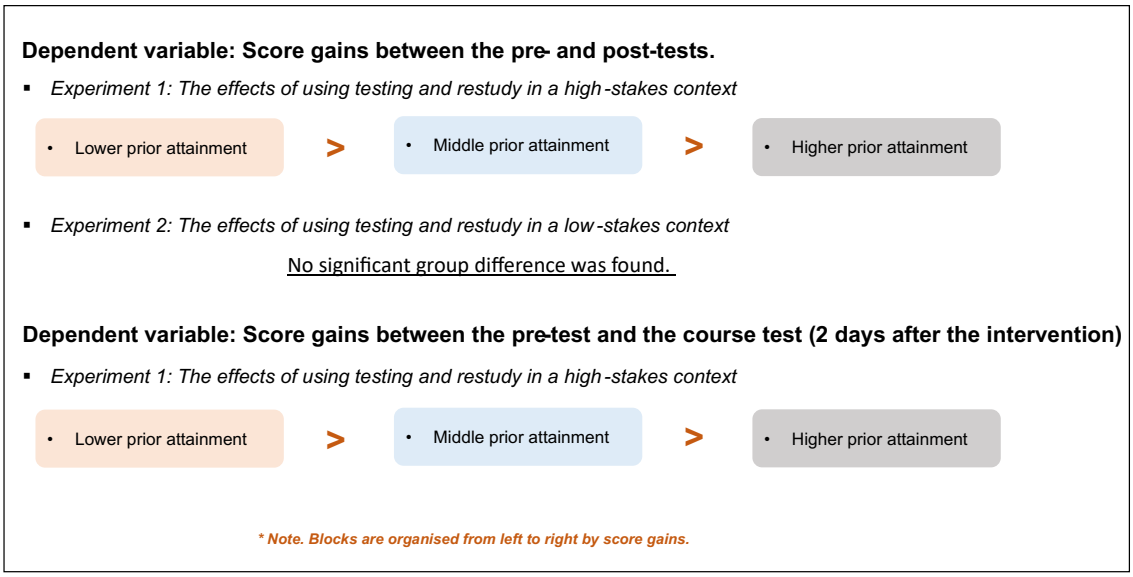


Figure 19 Block diagram of significant differences between different prior attainment levels on **score gains** in Experiments 1 and 2.

Score gains between the pre- and post-tests

The results of the three-way ANOVA indicated that there was a significant main effect of prior attainment level on students' score gains in between the pre- and post-tests, F (2, 263) = 18.89, $p < .001$, $\eta_p^2 = .126$. Post-hoc analysis showed that on average, students with lower prior attainment benefited more from the intervention than those with higher prior attainment (*Mean difference* = 11.30, *SE* = 1.85, $p < .001$) and those with middle attainment (*Mean difference* = 5.29, *SE* = 1.91, $p < .05$). The interaction between prior attainment levels and experiments was significant, F (2, 263) = 7.19, $p < .001$, $\eta_p^2 = .052$. Results of the pairwise comparisons revealed that only in Experiment 1, students with lower prior attainments obtained greater score gains in the post-test than students with middle attainments and high attainments (*p*s <. 001). Meanwhile, students with middle attainments also obtained greater score gains than those higher performers ($p < .01$). However, this pattern was not

observed in Experiment 2, as there was no difference on score gains across the three prior attainment levels in Experiment 2. The interaction between test preparation conditions and prior attainment levels was not significant, F (6, 263) = 0.78, p =.584, η_p^2 = .018, nor was the interaction between test preparation conditions, experiments, and prior attainment levels, F (6, 263) = 1.66, p =.130, η_p^2 = .037.

Score gains between the pre-test and the course test (Experiment 1 only)

Given that students in Experiment 2 took the final course test prior to taking part in the experiment, their performance in the final course test is not relevant to the investigation of the effect of test preparation from the intervention. Therefore, only data from Experiment 1 were used to undertake the analysis of whether students' prior attainment levels associated with score gains in final course test. A two-way ANOVA model was conducted, using test preparation conditions and students' prior attainment level as the two independent variables, and score gains in the final course as the dependent variable. Result revealed that prior attainment level had a main effect on score gains in the final course test, F (2, 160) = 44.08, p < .001, η_p^2 = .355. Pairwise comparison results suggested a similar pattern with score gains in the post-test; that students with lower prior attainment obtained greater score gains than students with middle and high prior attainments (*Mean difference* = 6.44 (*SE*=1.43) and 16.48 (*SE*=1.76) respectively, *p*s < .001), meanwhile students with middle prior attainment also obtained greater score gains than students with higher prior attainment (*Mean difference* = 10.03, *SE* = 1.78, p < .001). However, the interaction between test preparation conditions and prior attainment levels was not significant, F (6, 160) = 0.46, p = .836, η_p^2 = .017. The post-hoc tests did not reveal any significant paired differences (*p*s > .05).

Experiments 3 and 4

Regarding students in Experiments 3 and 4 (the cohort enrolled in the educational psychology course during 2020-2021), the relationships between students' absolute prior

attainment levels and the score gains in the post-test and in the final test were explored, as well as how the variable of prior attainment interacted with different test preparation conditions in the two experiments and the control condition. Two-way ANOVA models were performed on score gains in the post-test and the final test, using test preparation conditions and students' prior attainment levels as the two independent variables. Significant group differences revealed by ANOVA models are presented in Figure 20 and outlined below.

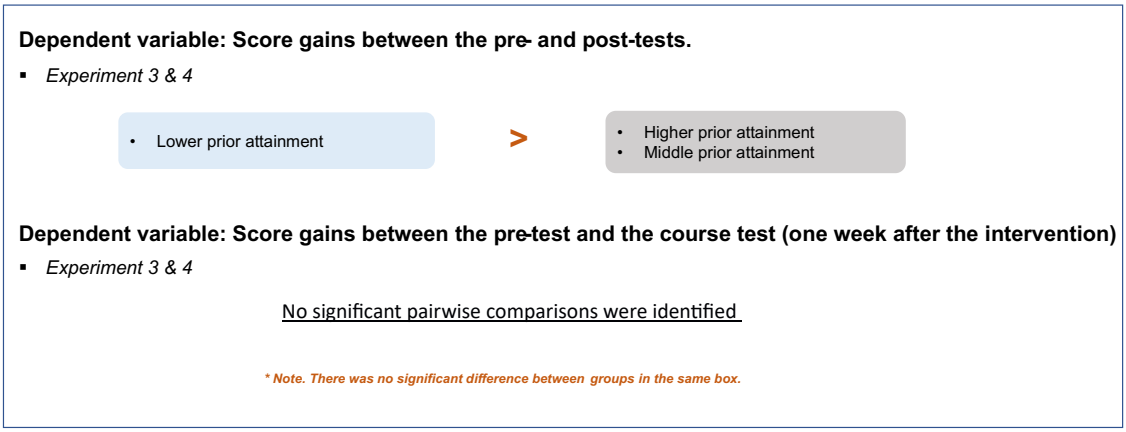


Figure 20 Block diagram of significant differences between different prior attainment levels on **score gains** in Experiments 3 and 4.

Score gains between the pre- and post-tests

The results indicated that there was a main effect of prior attainment level on post-test score gains, F (2, 391) = 7.88, $p < .001$, $\eta_p^2 = .039$. The follow-up pairwise comparisons showed that in this cohort, students who had lower prior attainments obtained greater score gains than students with medium prior attainments (*Mean difference* = 2.87, $p < .05$) and high prior attainments (*Mean difference* = 5.50, $p < .001$). The interaction between test preparation conditions and prior attainment levels was not significant, F (10, 391) = 0.66, $p = .763$, $\eta_p^2 = .017$. However, the post-hoc pairwise comparisons suggested that in *Restudy plus metacognition* (in Experiment 3) and the *Combined used of the testing and restudy* (the Experiment 4) groups, students with lower prior attainment levels benefited more than their

peers who had higher prior attainment level, *p*s < .05. These results suggested that there were significant differences between prior attainment levels within the two specific conditions, but the effect of prior attainment level on score gains did not significantly depend on what test preparation conditions were. Other pairs were not significantly different, *p*s > .05.

Score gains between the pre-test and the course test

Given that the Experiments 3 and 4 were concurrently conducted, one week prior to the final course test, the timing of the two experiments with the final course test was not used as an independent variable as it did for Experiments 1 and 2. Using score gains between the pre-test and the final course test as the dependent variable, absolute prior attainment level and test preparation conditions as the two independent variable, the two-way ANOVA model revealed that the main effect of prior attainment level was not significant, F (2, 391) = 0.60, p = .550, η_p^2 = .003. No significant pairwise comparisons were identified by the post-hoc Bonferroni tests. The interaction between the prior attainment level and the test preparation condition was not significant, F (10, 391) = 0.20, p = .996, η_p^2 = .005, and no significant pairwise comparisons were found in the post-hoc Bonferroni test, *p*s > .05.

To sum up, the analyses of the relationship between absolute prior attainment levels and score gains showed that there were significant differences between prior attainment levels on score gains in the post-test in Experiment 1, Experiment 3 and Experiment 4, which were high-stakes testing conditions. A similar pattern was found in these three experiments: the lower the prior attainment, the more the gains for the student. However, no significant group difference on score gains was found in Experiment 2. The different findings might relate to the stakes associated to the results of students' preparation. In the low-stakes setting in Experiment 2, no stake was associated with the outcomes of students' study following the intervention, thus they did not face the same pressure as those in the other three experiments. Therefore, students in Experiment 2 achieved generally low gains and the difference between prior attainment levels was diminished. However, due to the absolute way in which their

performances were classified, there were more high prior attainment performers and less low prior attainment performers in Experiment 2, and this might be the other reason of the small gaps in score gains across prior attainment levels. To test this possibility, students' score gains between the relative prior attainment levels (within experiment) were compared in the next section.

Regarding the score gains in course tests, a similar pattern was found in Experiment 1, with lower prior attainment performers benefiting more. However, this pattern was not observed in Experiments 3 and 4. A possible reason is that in Experiment 1, students took the course test two days after the experiment, while in Experiments 3 and 4, students took the course test one week later. With the increasing time interval, the effect of test preparation may not be retained.

7.4.2 Relative prior attainment levels and score gains

In the analyses above, students' prior attainment level was categorised across experiments using the same pre-test, thus some experiments had generally higher attainment due to the timing of the experiment or the way students were allocated. Specifically, students in Experiment 2 showed higher level of general attainment than students in Experiment 1 because they took part in the study later and had more familiarity with relevant course materials. Students in Experiment 4 also had higher level of attainment in general than students in Experiment 3 and those in the control condition, mostly likely because they were assigned to different conditions by non-random class-level allocation. A further examination was conducted to investigate whether the relative level of prior attainment, which was operationalised by comparing students with their peers in the same experiment, had a different relationship with score gains than the absolute level of prior attainment. In this case, within each experiment students were partitioned into three relative performance levels based on whether their pre-test scores fell within the upper, middle, or lower one third of all students who participated in the experiment. Table 25 presents score gains of students with

relative attainment levels in all conditions of the four experiments. Note that there are more equal numbers in the low, middle and high groups within each experiment's conditions, due to the relative way in which the classifications were conducted.

Table 25 Score gains among relative prior attainment levels across test preparation conditions in each experiment.

Experiment	Condition	Relative prior attainment	Score gain between pre- and post-tests	Score gain between the course tests and pre-test
Experiment 1	Testing & Feedback	Low (n=19)	22.47 (13.01)	30.32 (9.38)
		Middle (n=14)	4.57 (10.74)	22.86 (7.39)
		High (n=13)	-1.15 (12.51)	12.77 (10.07)
	Testing only	Low (n=13)	16.15 (11.97)	25.38 (7.92)
		Middle (n=16)	5.31 (14.03)	21.75 (7.90)
		High (n=11)	-4.09 (14.57)	11.55 (9.46)
	Restudy	Low (n=12)	12.33 (12.99)	26.50 (8.67)
		Middle (n=16)	7.50 (11.17)	22.69 (6.85)
		High (n=13)	0.15 (13.64)	14.23 (7.49)
	Control	Low (n=13)	4.92 (6.38)	23.31 (8.54)
		Middle (n=14)	-1.21 (7.55)	19.86 (7.22)
		High (n=18)	-3.94 (7.45)	10.61 (9.79)
Experiment 2	Testing & Feedback	Low (n=10)	11.40 (11.47)	
		Middle (n=10)	5.80 (11.12)	
		High (n=12)	0.92 (9.10)	
	Testing only	Low (n=8)	0.63 (10.80)	
		Middle (n=13)	8.08 (9.12)	
		High (n=7)	0.00 (8.17)	
	Restudy	Low (n=9)	8.44 (11.22)	
		Middle (n=10)	11.30 (13.46)	
		High (n=12)	2.08 (10.85)	
	Control	Low (n=11)	8.55 (15.14)	
		Middle (n=6)	-0.17 (6.27)	
		High (n=7)	-3.57 (5.44)	

Experiment 3	Testing & Feedback	Low (n=12)	8.92 (8.95)	31.83 (10.98)
		Middle (n=18)	11.28 (8.82)	31.50 (8.54)
		High (n=9)	4.89 (9.91)	30.78 (6.63)
	Testing & Feedback plus Metacognition	Low (n=11)	13.36 (8.18)	28.55 (11.41)
		Middle (n=15)	8.33 (10.79)	28.20 (10.27)
		High (n=12)	7.67 (8.70)	26.67 (6.88)
	Restudy	Low (n=15)	10.13 (6.51)	29.13 (14.96)
		Middle (n=14)	7.36 (8.72)	23.21 (13.76)
		High (n=10)	7.70 (9.97)	28.00 (10.04)
	Restudy plus Metacognition	Low (n=13)	13.92 (5.92)	28.15 (9.22)
		Middle (n=13)	10.77 (8.78)	28.77 (8.76)
		High (n=11)	4.36 (6.27)	25.91 (9.30)
	Control	Low (n=35)	3.69 (6.52)	20.89 (8.91)
		Middle (n=24)	1.21 (8.06)	23.04 (13.76)
		High (n=45)	-3.36 (9.49)	20.44 (10.42)
Experiment 4	Combined use of testing and restudy	Low (n=46)	10.83 (10.01)	24.35 (11.99)
		Middle (n=56)	8.07 (11.07)	22.46 (10.49)
		High (n=50)	5.68 (8.60)	22.54 (12.03)
	Control	Low (n=40)	3.33 (7.21)	21.30 (9.67)
		Middle (n=29)	-0.59 (7.45)	22.14 (12.55)
		High (n=35)	-3.11 (10.22)	20.29 (10.61)

A series of two-way ANOVA models were conducted using data from each experiment, with the score gains of post-tests and course tests as dependent variables, the test preparation condition and the relative prior attainment level as two independent variables. Significant group differences identified by two-way ANOVA models on score gains of the post-test and score gains of the course are shown in Figures 21 and 22.

Score gains between the pre- and post-tests.

Based on data from Experiment 1, the two-way ANOVA results indicated that there was a significant main effect of relative prior attainment level, F (2, 160) = 27.32, $p < .001$, η_p^2 = .243. The following post-hoc Bonferroni tests suggested that students with relative lower prior attainment achieved a significantly greater score gain than those with middle (*Mean*

Difference= 11.03 *SE* = 2.17, $p < .001$) and higher (*Mean Difference*= 16.67, *SE* = 2.18, $p < .001$) prior attainment. Students with middle prior attainment also achieved, on average, significantly larger score gains than the high-performing students (*Mean Difference*= 5.63, *SE* = 2.17, $p < .05$). The interaction between the prior attainment and test preparation group was not significant, F (6, 160) = 1.56, $p = .163$, $\eta_p^2 = .055$. No significant pairwise comparisons were identified by the post-hoc tests.

Performing a similar two-way ANOVA using data from Experiment 2, the results revealed that there was a main effect of prior attainment level on students' score gains, F (2, 103) = 4.96, $p < .01$, $\eta_p^2 = .088$. The Bonferroni post hoc tests indicated that students with relatively low prior attainment achieved an averagely larger score gain than those with higher attainment (*Mean Difference*= 7.32 *SE* = 2.48, $p < .05$). Students with relatively middle prior attainment also had a significantly greater score gain (*Mean Difference*= 6.76, *SE* = 2.46, $p < .05$) than those with higher prior attainment, $p < .05$. The interaction between test preparation group and relative prior attainment level on the score gains between pre-test and post-test was not significant, F (6, 103) = 1.14, $p = .347$, $\eta_p^2 = .062$. No significant pairwise comparisons were identified by the post-hoc tests.

The two-way ANOVA models were also conducted on data from Experiment 3, using score gains in the post-test and the course test as the dependent variable respectively, prior attainment level and test preparation groups as the two independent variables. Regarding differences in score gains of the post-test, similar findings with Experiment 1 were found. The results indicated that a significant main effect of relative prior attainment level was identified, F (2, 242) = 7.77, $p < .001$, $\eta_p^2 = .060$. The post hoc analysis showed that relatively low-performing students achieved a significantly greater score gain than high prior attainment performers (*Mean Difference*= 5.75, *SE* = 1.47, $p < .001$) in the post-test. The relatively middle prior attainment performers also achieved a significantly greater score gain than those with relative higher prior attainment (*Mean Difference*= 3.54, *SE* = 1.43, $p <. 05$).

The interaction between test preparation groups and attainment level was not significant, F (8, 242) = 0.79, p = .614, η_p^2 = .025. No significant pairwise comparisons between groups of the two variables was found in post-hoc Bonferroni tests.

Regarding score gains in the post-test in Experiment 4, the two-way ANOVA found that the main effect of prior attainment level was significant when score gains in the post-test was used as the dependent variable, F (2, 249) = 8.07, p < .001, η_p^2 = .061. The post hoc tests revealed that students with relatively low prior attainment achieved significantly greater score gain in the post-test compared with those with high prior attainment (*Mean Difference*= 5.28, *SE* = 1.44, p < .001), while no significant difference were found between students with relatively middle prior attainment and those with low and high prior attainment (*p*s > .05). The interaction between group and prior attainment level was not significant, F (2, 249) = 0.12, p = .887, η_p^2 = .001.

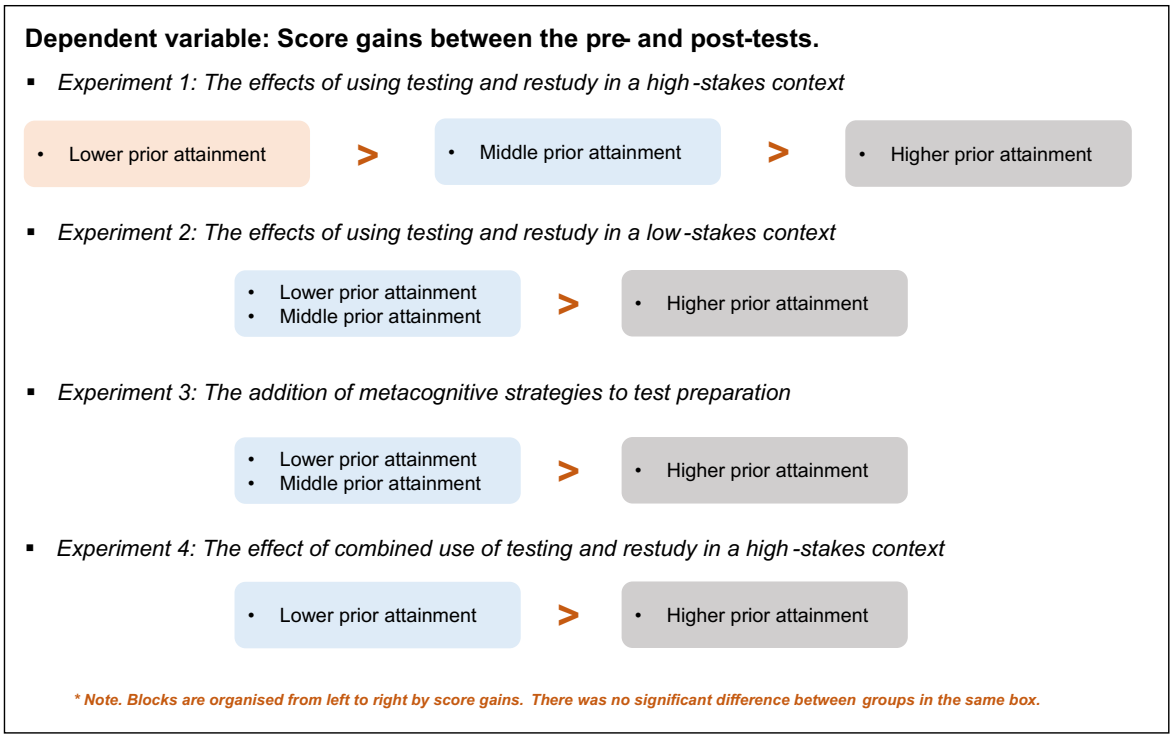


Figure 21 Block diagram of significant group differences between different relative prior attainment levels on **score gains on post-tests**.

Score gains between the pre-test and the course test

In terms of the score gains in course tests in Experiment 1, similar patterns were found. The two-way ANOVA model showed that only the main effect of prior attainment level was significant, F (2, 160) = 40.01, $p < .001$, $\eta_p^2 = .333$, in that students with relative lower prior attainment level benefited significantly more than those with middle prior attainment level (*Mean Difference*= 4.74, *SE* = 1.57, $p < .01$) and higher attainment (*Mean Difference*= 14.53, *SE* = 1.58, $p < .001$). Meanwhile students with middle prior attainment benefited significantly more than students with high prior attainment (*Mean Difference*= 9.79, *SE* = 1.57, $p < .001$). The interaction between relative prior attainment level and test preparation group, however, were not significant, F (6, 160) = 0.34, $p = .914$, $\eta_p^2 = .013$. The post-hoc tests did not reveal any significant pairwise comparisons between the groups of the two variables.

Using data from Experiment 3, the two-way ANOVA indicated that the main effect of relative prior attainment was not significant, F (2, 242) = 0.27, $p = .762$, $\eta_p^2 = .002$. The interaction between group and relative prior attainment level was also not significant, F (8, 242) = 0.44, $p = .895$, $\eta_p^2 = .014$.

In Experiment 4, the two-way ANOVA showed that the main effect of relative prior attainment level on score gains in the course test was not significant, F (2, 249) = 0.34, $p = .714$, $\eta_p^2 = .003$. And the interaction between group and relative prior attainment level was not significant, F (2, 249) = 0.31, $p = .734$, $\eta_p^2 = .002$. The post-hoc tests did not identify any significant pairwise comparison.

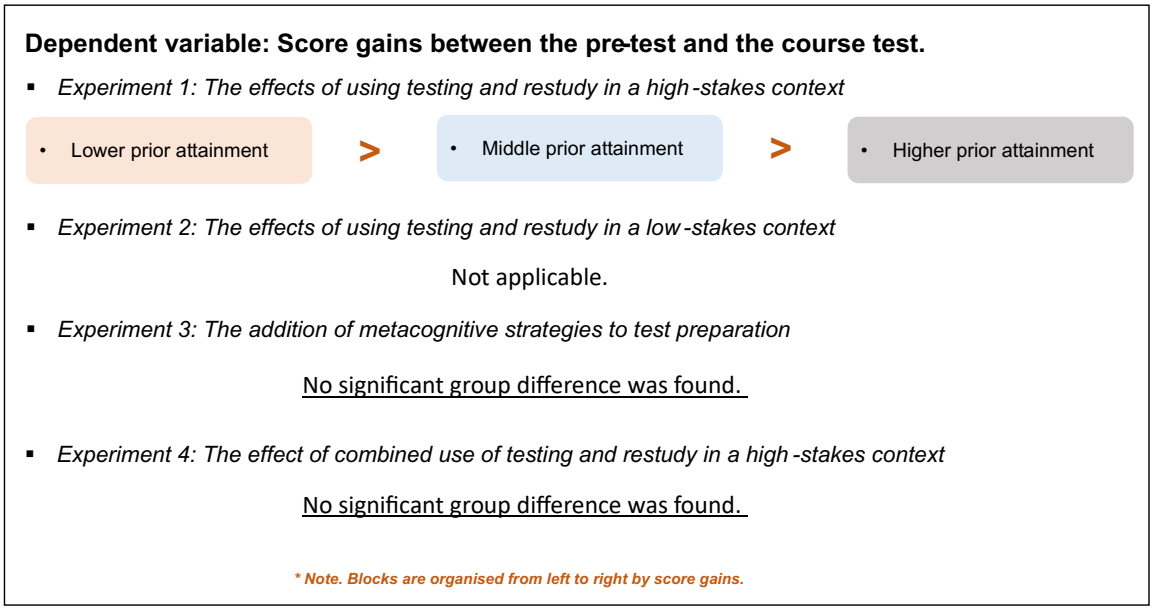


Figure 22 Block diagram of significant group differences between different relative prior attainment levels on **score gains on final course tests**.

To summarise, the above findings on the relationship between the prior attainment level (in both *absolute* and *relative* senses) and score gains revealed that in general students with lower prior attainment levels achieved more gains in post-tests than those with higher attainment levels on score gains. It is worth noting that the analyses on *absolute* prior attainment and *relative* prior attainment had one different finding. Specifically, that the former analysis found no significant difference on post-test score gains across *absolute* prior attainment levels in Experiment 2. In contrast, the latter analysis found that relatively lower and middle prior performers received greater gains in the post-test than relatively high performers in Experiment 2. This pattern was also identified in Experiments 1 and 3 in terms of the analyses of *relative* prior attainment. A possible reason for the difference might be that under the *absolute* prior attainment index, more students in Experiment 2 were categorised as with higher prior attainment levels, thus the difference between the three categories of students might be reduced in this experiment. But when compared with their peers who

participated in the experiment at the same time, relative lower performers in Experiment 2 still improved more than relative higher performers.

As for gains in the course test taken later, both analyses of the absolute and relative prior attainment found that only in Experiment 1, the lower prior attainment the students had, the more gains they achieved in the final course test after the intervention. Similar results were not observed in Experiment 3 and Experiment 4. One possible reason might be associated with the time interval between the intervention and the final course test. In Experiment 1, students took the course test two days after participating in the experiment, while in Experiment 3 and 4, the course test was undertaken one week after the experiment. It could be that as the time interval between the intervention and the test increased, the gains that students with lower prior attainment level received from the interventions may be eroded, or students with higher prior attainment level may engage in more effective self-study sessions before the final course test and thus raise their score gains in the course test. However, as there was no manipulation between the intervention and the course test, the speculations cannot be verified with direct information. In the following analysis, I attempt to investigate the differences in study behaviours and levels of monitoring accuracy of students with different levels of prior knowledge to explore whether students' score gains associated with specific study behaviours or the improved monitoring accuracy during the test preparation process.

7.5 Prior attainment levels, test preparation behaviours and monitoring accuracy

The analyses presented in the above section revealed that in general, students' prior attainment related to the magnitude of test preparation effects. As such, it would be informative to explore whether students with different levels of prior attainment showed different test preparation behaviours and different levels of monitoring accuracy, which might explain the variance in the effectiveness of test preparation. In these (quasi-) experiments,

students' study behaviours were also reflected in the amount of time they spent in study, the number of overall study attempts they used, and the number of attempts using every specific strategy (testing or restudy) on every-to-be-reviewed topic. Data on time use during the intervention were recorded in all four (quasi-) experimental studies, while study attempts were recorded in Experiment 3 and 4. Data on students' monitoring accuracy were also recorded in Experiments 3 and 4. The prior attainment level in the absolute sense was used in the analyses, as it enabled meaningful comparisons across experiments and was found to show stable findings across experiments in the analyses reported above. A series of ANOVA models were performed to examine the group differences on time use and study attempts between prior attainment levels and test preparation conditions across experiments. Significant group differences are summarised in Figures 23 and 24.

7.5.1 Time use during the interventions.

First, a three-way ANOVA model was conducted using data from Experiments 1 and 2, with time use as the dependent variable, prior attainment levels, test preparation conditions, and experiment sequence as the three independent variables. The results showed that there was no main effect of prior attainment level on time use, F (2, 200) = 0.72, p = .487, η_p^2 = .007. Post-hoc tests confirmed that the three levels of prior attainment groups spent similar time in the self-paced test preparation intervention. Interactions between prior attainment levels and test preparation conditions, and between prior attainment levels and the experiment sequence were not significant, F (4, 200) = 1.09, p = .365, η_p^2 = .021; and F (2, 200) = 0.54, p = .582, η_p^2 = .005. Similarly, the interaction between the three independent variables was not significant, F (4, 200) = 0.62, p = .651, η_p^2 = .012. No significant pairwise comparison between prior attainment levels and other factors was found,

Based on data from Experiment 3 and Experiment 4, a two-way ANOVA was performed, with time use as the dependent variable, test preparation conditions and prior attainment levels as two independent variables. The results suggested that there was no

significant effect of students' prior attainment level on time uses during the intervention, F (2, 290) = 2.09, p = .126, η_p^2 = .014. However, the post-hoc Bonferroni test revealed that there were significant differences between different prior attainment levels on time uses. Students with higher prior attainment spent averagely more time on the intervention than students with low (*Mean difference* = 3.76, *SE* = 0.97, $p < .001$) and middle (*Mean difference* = 2.27, *SE* = 0.93, $p < .05$) prior attainment. Besides, the interaction between prior attainment levels and test preparation conditions was not significant, F (8, 290) = 0.857, p = .553, η_p^2 = .023. However, the results of pairwise comparisons revealed that for students with low prior attainment, those in the *Combined use of strategies* condition spent averagely more time than those in the *Restudy* (*Mean difference* = 13.21, *SE* = 2.03, $p < .001$), *Testing with feedback* **(***Mean difference* = 10.99, *SE* = 2.14, $p < .001$**)**, and *Restudy plus metacognition* (*Mean difference* = 10.30, *SE* = 2.08, $p < .001$) conditions. Similar patterns were found for students with middle and high prior attainment levels, that those in the *Combined use of strategies* condition spent averagely more time than those in the *Restudy* **(***Mean difference* = 12.48, *SE* = 2.03, $p < .001$; *Mean difference* = 14.27, *SE* = 2.29, $p < .001$), *Testing with feedback* **(***Mean difference* = 5.99, *SE* = $p < .05$; *Mean difference* = 9.27, *SE* = 2.68, $p < .01$, respectively**),** *and Restudy plus metacognition* **(***Mean difference* = 6.74, *SE* = 1.89, $p < .01$; *Mean difference* = 7.35, *SE* = 2.40, $p < .05$, respectively) conditions. Within every test preparation condition, no significant difference between prior attainment levels was found, ps > .05.

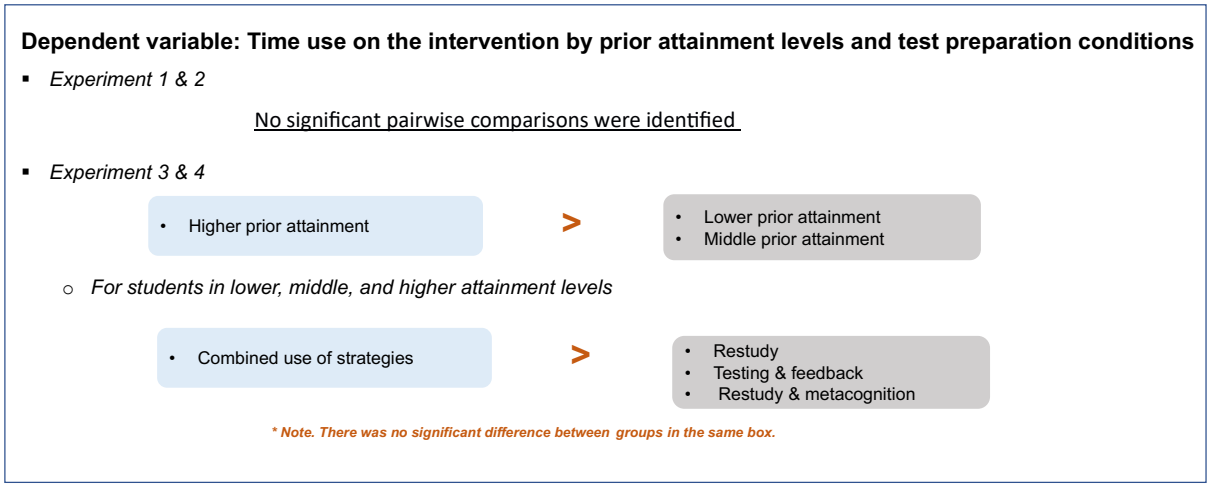


Figure 23 Block diagram of significant group differences between different prior attainment levels and test preparation conditions on **time use during the intervention**.

7.5.2 Overall study attempts during the intervention (Experiments 3 and 4 only).

The relationship between students' overall study attempts and their prior attainment levels was investigated based on data from Experiments 3 and 4. A two-way ANOVA model was conducted, using overall study attempts as the dependent variable, prior attainment level and test preparation condition as the two independent variables. Significant group differences revealed by the ANOVA model are summarised in Figure 24. Results revealed that there was a significant main effect of test preparation group on overall study attempts, F (4, 290) = 53.90, $p < .001$, $\eta_p^2 = .426$. The post-hoc tests indicated that the *Combined use of testing and restudy* condition (the Experiment 4) had significantly more study attempts than other four conditions in Experiment 3, including the *Testing with feedback plus metacognition* (Mean difference = 3.34, $SE = .729$, $p < .001$), *Restudy plus metacognition* (Mean difference = 6.54, $SE = .737$, $p < .001$), *Testing with feedback* (Mean difference = 7.61, $SE = .721$, $p < .001$), and *Restudy* (Mean difference = 7.81, $SE = .721$, $p < .001$). Meanwhile, the *Testing with feedback plus metacognition* condition also had significantly more study attempt than the *Restudy plus metacognition* (*Mean difference* = 3.20, $SE = .729$, $p < .01$), *Testing with feedback* (*Mean difference* = 4.27, $SE = .916$, $p < .001$), and *Restudy* (*Mean difference* = 4.48,

SE = .916, $p < .001$). The main effect of prior attainment level on overall study attempts was not significant, F (2, 290) = 0.04, p =.961, η_p^2 = .000. The post hoc tests did not find any significant comparisons between the three prior attainment groups. The interaction between test preparation conditions and prior attainment levels was not significant, F (8, 290) = 1.00, p =.434, η_p^2 = .027. However, as can be seen in Table 26, post-hoc analyses suggested that students with lower prior attainment had significantly more study attempts when they were in the *Combined use of strategies* condition than other four conditions, ps < .001; whereas students with middle and high prior attainment in *Combined use of strategies* condition used significantly more study attempts than those in conditions except for those in *Testing with feedback plus metacognition***.** Additionally, students with middle prior attainment in *Testing with feedback plus metacognition* condition also had more study attempts than in other three conditions except for those in *Combined use of strategies*. Within the *Combined use of strategies* condition, students with low prior attainment had significantly more study attempts than those with high prior attainment (*Mean difference* = 2.59, *SE* = .812, $p < .01$).

Table 26 Significant difference in study attempts between test preparation conditions within prior attainment levels.

Prior attainment level	Test preparation condition (a)	Test preparation condition (b)	Mean difference (a-b)	Std. Error	p	95% CI-Low	95% CI-High
Low	Combined use of testing and restudy	Testing with feedback plus metacognition	5.83*	1.365	<.001	1.97	9.69
		Testing with feedback	9.41*	1.279	<.001	5.79	13.03
		Restudy plus metacognition	8.28*	1.244	<.001	4.76	11.79
		Restudy	9.56*	1.213	<.001	6.13	12.99
Middle	Combined use of testing and restudy	Testing with feedback plus metacognition	2.44	1.128	.312	-.75	5.63

		Testing with feedback	7.30*	1.083	<.001	4.23	10.36
		Restudy plus metacognition	6.63*	1.215	<.001	3.19	10.07
		Restudy	7.49*	1.215	<.001	4.05	10.93
	Testing with feedback plus metacognition	Testing with feedback	4.85*	1.342	.003	1.06	8.65
		Restudy plus metacognition	4.19*	1.450	.042	0.09	8.29
		Restudy	5.05*	1.450	.006	0.94	9.15
High	Combined use of testing and restudy	Testing with feedback plus metacognition	2.57	1.371	.621	-1.31	6.45
		Testing with feedback	6.54*	1.604	<.001	2.00	11.08
		Restudy plus metacognition	4.97*	1.435	.006	0.91	9.03
		Restudy	6.87*	1.371	<.001	2.99	10.75

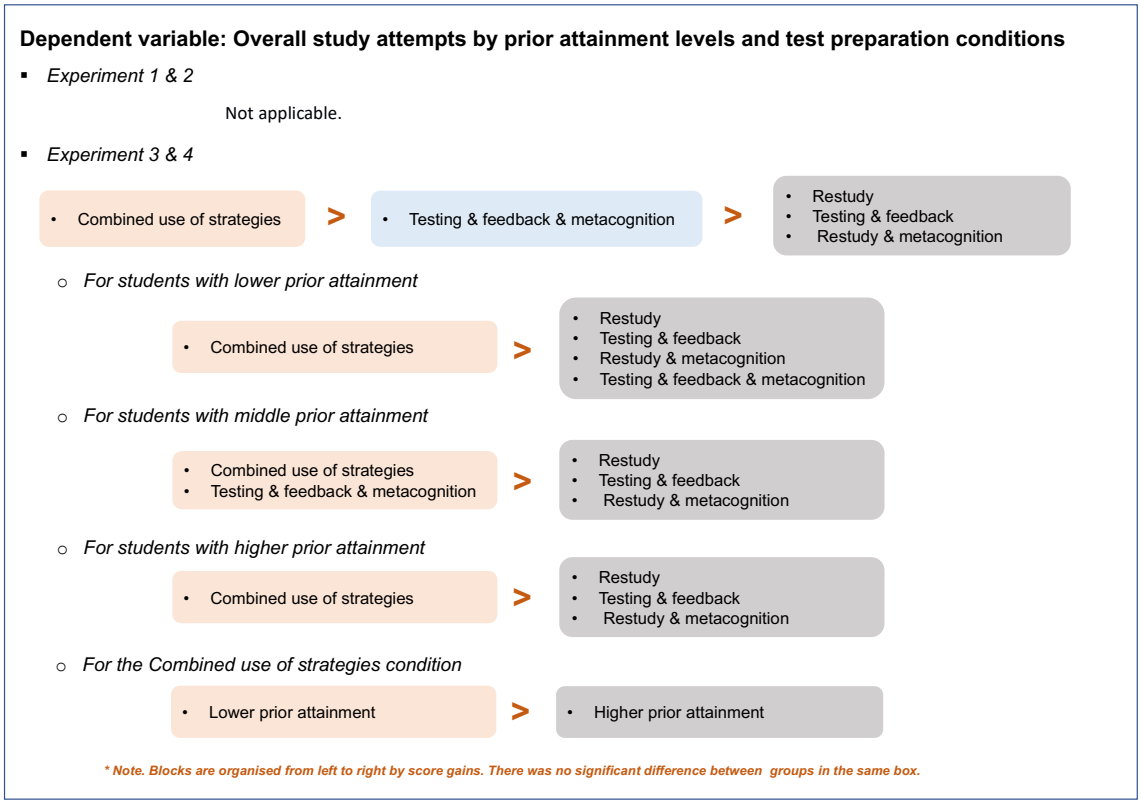


Figure 24 Significant differences between prior attainment levels and test preparation conditions on **study attempts.**

7.5.3 Testing and restudy attempts during the intervention (Experiment 4 only).

To further understand whether students with different prior attainment levels have different propensities to apply particular strategies, one-way ANOVA models were employed based on data from Experiment 4. In this experiment (i.e., the *Combined use of strategies* condition), students were allowed to choose to use restudy or testing to study educational psychology topics as many times as they want. Two one-way ANOVA models were used to compare the overall numbers of testing attempts and rereading attempts across different prior attainment levels. Significant group differences are shown in Figure 23. The results of ANOVA on testing attempts showed that there was no significant difference on students' use of testing attempts between different prior attainment levels, F (2, 149) = 0.16, p = .851, η_p^2 = .049, and no significant pairwise comparison was found. Regarding students' use of restudy, a significant difference was identified across different prior attainment levels, F (2, 149) = 4.58, $p < .05$, η_p^2 = .058. Post-hoc tests revealed that students with low prior attainment utilised the restudy strategy (*Mean* = 11.81, *SD* = 5.32) more frequently than students with high prior attainment (*Mean* = 8.74, *SD* = 4.62), $p < .01$. No other significant difference on the frequency of restudy attempts was observed.

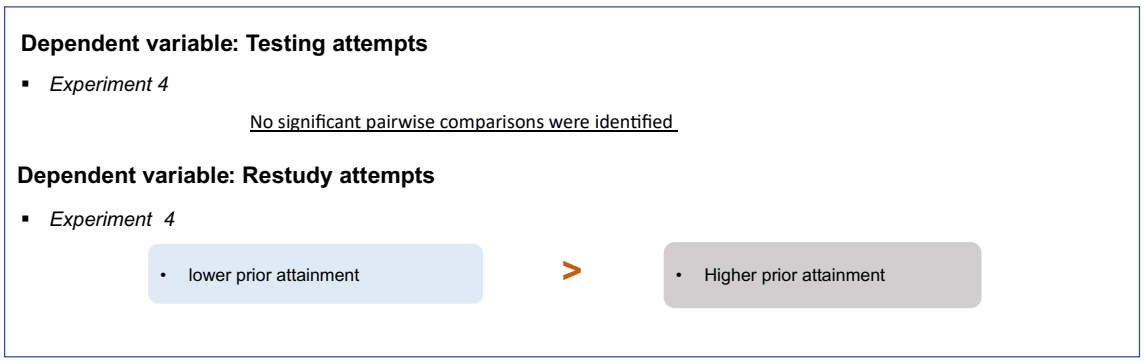


Figure 25 Significant differences between prior attainment levels and test preparation conditions on **testing and restudy attempts.**

7.5.4 Absolute accuracy of monitoring (Experiments 3 and 4).

Previous research indicated that students with different prior attainment levels had different levels of accuracy, with low-performing students usually being less accurate on the judgement of their test performance (Bol & Hacker, 2001; Carpenter et al., 2016). It is assumed that high monitoring accuracy could help students effectively regulate their study and thus improve the test performance (Nelson & Narens, 1990; Winne, 1996). Based on this assumption, one possible explanation of why lower prior attainment benefited more from the intervention is that their monitoring accuracy improved and thus effectively regulated their test preparation behaviours. To testify this explanation, the relationship between prior attainment and absolute accuracy of monitoring was investigated. The description information of absolute accuracy is presented in Table 27.

Table 27 Absolute monitoring accuracy of the pre-test, post-test and changes.

Experiment	Group	Prior attainment	N	Pre-test		Post-test		Changes	
				Mean	SD	Mean	SD	Mean	SD
Experiment 3	Testing with feedback	Low	13	0.17	0.05	0.18	0.07	0.01	0.08
		Middle	19	0.15	0.03	0.12	0.05	-0.03	0.07
		High	7	0.13	0.06	0.11	0.04	-0.01	0.06
	Testing with feedback plus metacognitive strategies	Low	11	0.16	0.03	0.16	0.08	-0.01	0.08
		Middle	17	0.15	0.05	0.14	0.05	-0.01	0.06
		High	10	0.13	0.03	0.13	0.05	-0.003	0.06
	Restudy	Low	14	0.14	0.04	0.14	0.06	-0.01	0.07
		Middle	15	0.16	0.04	0.13	0.05	-0.03	0.07
		High	10	0.12	0.06	0.13	0.05	0.01	0.06
	Restudy plus metacognitive strategies	Low	14	0.15	0.05	0.14	0.05	-0.01	0.05
		Middle	14	0.13	0.04	0.12	0.04	-0.01	0.05
		High	9	0.12	0.05	0.16	0.05	0.03	0.03
	Total		153	.145	.046	.138	.054	-.007	.063
Experiment 4		Low	41	0.16	0.05	0.14	0.04	-0.02	0.06
		Middle	50	0.14	0.04	0.14	0.04	-0.003	0.07

	Combined use of testing and restudy	High	61	0.14	0.05	0.13	0.05	-0.004	0.07
	Total		152	0.15	0.05	0.14	0.05	-0.01	0.07
Control	Control	Low	43	0.15	0.06	0.14	0.05	-0.004	0.07
		Middle	24	0.13	0.03	0.15	0.04	0.02	0.05
		High	37	0.12	0.04	0.15	0.05	0.02	0.05
	Total		104	0.13	0.05	0.15	0.05	0.01	0.06

The two-way ANCOVA model was performed, with students' absolute accuracy of monitoring on the post-test as the dependent variable, prior attainment levels and test preparation conditions as the independent variable, and absolute accuracy of monitoring on the pre-test as the covariate. The results indicated that there was no significant main effect of prior attainment levels, F (2, 390) = 2.09, p = .125, η_p^2 = .011. Similarly, there was no significant main effect of test preparation conditions, F (5, 390) = 0.67, p = .648, η_p^2 = .008. The interaction between prior attainment levels and test preparation conditions was not significant, F (10, 390) = 1.34, p = .206, η_p^2 = .033. No significant group difference was found in the post-hoc tests, ps > .05.

To summarise, the analyses presented in this section revealed that in general the *Combined use of strategies* condition (i.e., Experiment 4) spent on average more time on test preparation and used more study attempts. This finding is not surprising because students in this condition were reminded that they could choose different strategies to study different topics or study a topic for multiple times using different strategies. As discussed in Paper 3, both the quantitative and qualitative data confirmed that students to a large extent blended the use of different strategies as they believed this could maximise the effect of their study.

Findings in previous sections suggested that in general students with lower prior attainment had greater score gains in post-tests, and the analyses on study behaviours in this

section showed that they tended to use more study attempts and use restudy more frequently in Experiment 4. Three plausible assumptions can be deduced: 1) studying more times can produce more gains; 2) more restudy practices can produce more gains; and 3) more overall attempts and more restudy attempts are reflections of effective regulation of study, thereby producing more gains. In terms of the first assumption, in general students in Experiment 4 showed more study attempts than students in Experiment 3, even for the two metacognition groups where similar monitoring and control tests were embedded as Experiment 4 did. The analysis of effects of the intervention in Experiment 4 presented in both Section 7.2 and 7.3 suggested that, however, Experiment 4 did not have greater effect sizes, and leaving students to choose strategies for test preparation by their own seems even weaken the effect (though did not reach the significant level in this research). Thus, the first assumption is not supported by other findings in this research. For further verification, a Pearson correlation between the number of students' overall attempts and their score gains between the pre- and post-test in Experiment 4 was computed, and no significant correlation was found, $r(152) = 0.08$, $p = .337$. Similarly, the second assumption cannot be confirmed because Paper 2 did not reveal any superior effect of restudy practices. Additionally, there was no significant correlation between the number of restudy attempts and score gains, $r(152) = 0.13$, $p = .104$. Regarding the third assumption, the analysis showed that students with lower prior attainment did not improve their monitoring accuracy significantly more than others, thus the claim that they regulated their study more effectively cannot be supported.

Moreover, it is worth mentioning that findings also indicated that time spent on test preparation did not associate with score gains. The analyses showed that in Experiments 3 and 4, students with higher prior attainment averagely spent more time on every study attempt, as they used significantly more time and had less study attempts. However, they did not benefit more than middle and low prior attainment performers, as revealed by findings in previous sections. These findings implied that devoting more time on study did not

necessarily lead to more gains in this intervention. This may be because interventions in this research mainly provided factual knowledge on psychology topics, and students with higher prior attainment had already possessed a greater degree of knowledge mastery. Consequently, high-performing students did not achieve as much improvement as students with lower prior attainment who benefited more from building a fundamental understanding of the subject matter through learning factual knowledge during the interventions.

Therefore, it seems that the greater progress of lower performers may not relate to their studying more or using restudy more often. Lower-performing students may have gaps in their foundational knowledge, which hindered their performance in tests. Therefore, interventions focused on factual knowledge can be particularly beneficial for these students, even if they do not necessarily spend more time studying. This also explains why students with lower prior attainment chose to do restudy more often—probably because their knowledge mastery was low and testing was demanding for them, as suggested by the qualitative data in Paper 3. The relationship between prior attainment level and score gains are further discussed in Chapter 8 using relevant theories.

7.6 Overview and discussion of research findings

In this section, findings from the three papers and four sets of additional analyses are overviewed and discussed to build an integrated understanding of test preparation effects as well as students' strategy uses in a specific test preparation context in China.

7.6.1 Overall effect of test preparation on educational tests

The meta-analysis in Paper 1 reported a significant and positive test preparation effect on large-scale educational tests ($g = 0.38$, $p < .001$). Several moderators associated with the magnitude of effect size were identified, including the breadth of test preparation intervention (i.e., narrow and specific intervention had greater effect than broad intervention), workbooks (the use of workbooks led to greater effect sizes), test preparation management skills and test-

taking skills (i.e., instructions on the two skills yielded greater effect sizes). Taking practice tests and using sample items during test preparation were not significant moderators.

In subsequent (quasi-) experiments, the effect of two commonly used test preparation strategies on students' performance of a psychology course test were examined. The overall effect of the test preparation interventions among the four experiments was analysed by a continuously cumulating meta-analysis (CCMA), as well as how specific factors related to the test preparation practice and the context of test preparation moderated the effect. The CCMA identified a similar effect size with the meta-analysis for large-scale test, $g = 0.41$, $se = 0.19$, $p < .05$, and only the time interval between the intervention and the test was the significant moderator. Test preparation interventions had a significantly greater effect when the test was administered immediately after the intervention ($g = 0.60$, $se = 0.16$, $p < .01$) than the effect on delayed course test ($g = 0.21$, $se = 0.18$, $p = .259$), $b = 0.40$, $se = 0.11$, $p < .01$. Other factors (i.e., test preparation strategy, the use of metacognitive strategies, and stakes of test preparation contexts) did not have significant moderating effects on the magnitude of the effect size, though the sub-group analysis identified that only some specific conditions associated with significant test preparation effects (e.g., the high-stakes context associated with a significant effect), while others did not yield significant effects (e.g., the low-stakes context associated with a small and insignificant effect). Details are provided in the next section.

7.6.2 Factors associated with test preparation effects

The meta-analysis in Paper 1 suggested that subject domain might affect the effects of test preparation interventions, as it found that interventions on EFL had greater effects than interventions on verbal. Additionally, students' experience of test preparation might also be a possible influential factor, as students who had previous experience on test-taking and test preparation benefited more from interventions than those who did not have any previous exposure, based on the sub-group analyses of the meta-analysis. The two factors were not

subsequently explored by (quasi-) experiments in this research, and future investigation might provide more conclusive insights.

The four (quasi-) experiments in this research compared the effect of three test preparation practices (i.e., the use of testing, the use of restudy, and the combined use of both testing and restudy) with the control condition where students did not receive test preparation interventions. Findings from the four (quasi-) experiments seemed to be coherent with the findings on testing effect identified in Paper 1. The results suggested that in general testing with feedback was an effective test preparation strategy and it showed significant effects on both immediate post-tests and delayed course tests. However, the effect of testing did not exceed the effect of restudy, which suggested that in test preparation contexts where knowledge transfer is required, testing and restudy might have similar effects. Integrating this result and the findings on practice testing in Paper 1 leads to a possible inference: in test preparation contexts where knowledge transfer is required, testing is effective, but its effect might not be superior to other strategies.

The sole use of restudy had significant effects on the immediate post-test, but the effect might not be sustained because no significant effect of restudy practices was found on delayed course tests compared with the control. However, findings of Experiment 3 revealed that embedding metacognitive training and monitoring and control tasks during restudy practices could lead to significant effects on both immediate and delayed tests. This finding suggested, though not conclusively, the monitoring and control of restudy practices has greater potential to improve students' test performance and long-term learning. More future research is needed to justify whether longer training and more practices on metacognition can increase the effect of restudy practices and testing practices. After all, students' test preparation in real life can be extensive.

In addition, the stakes of the test seemed to relate to the effect of test preparation interventions. In low-stakes settings, test preparation did not have a significant effect, no matter what strategies were used.

Students' prior attainment levels is also one of the important factors that might affect the effect of test preparation interventions. A general pattern identified by additional analyses was that relatively, students with previously low-performing levels benefited more from test preparation interventions than those with higher performing levels. This pattern has been identified both within and across experiments. However, when it comes to whether students at a certain level of prior achievement benefited more from a specific test preparation strategy, the evidence provided by this research was ambiguous and did not show a clear pattern, as evidenced by the findings that prior attainment level did not interact significantly with test preparation conditions across experiments.

7.6.3 Test preparation behaviours and students' monitoring

The levels of students' prior attainment also related to students' study behaviours. The findings suggested that students with low prior attainment undertook significantly more study attempts in total and used restudy more frequently when they could choose which strategies to use.

In terms of study behaviours during the test preparation intervention, more findings were revealed by Paper 2 and Paper 3. Paper 2 reported that students in metacognition groups showed more study attempts than the regular groups (i.e., testing with feedback only and restudy only), and students in testing groups showed more study attempts than the two restudy groups. Paper 3 suggested that Chinese students in the study most frequently chose one-time testing to study a topic, followed by the combined use of testing and restudy, while repeatedly restudying a topic was the least frequent practice.

Students' study behaviours related to their confidence in the topic and the complexity of the tasks. Paper 2 found that students in the testing with feedback plus metacognition group

had a significantly greater confidence gain than those in the restudy plus metacognition group, indicating that testing practice might associated with greater confidence. This inference has been further confirmed by Paper 3. Both quantitative and qualitative data presented in Paper 3 indicated that students' confidence in a topic and how the topic was tested were two significant predictors of students' strategy choices. By examining students' monitoring accuracy in pre- and post-tests based on data from Experiment 3 and 4, the additional findings further suggested that test preparation interventions, no matter what specific practice was manipulated, did not improve the accuracy of students' metacognitive calibration. In general, students tended to be more over-confident after participating in the experiments.

To sum up, this research programme has provided rich synthetical and empirical findings around test preparation effect as well as students' use of testing and restudy for test preparation purposes. How these findings can be further integrated and explained based on relevant theories, and how they possibly contribute to the literature as well as their practice implications are discussed in the next chapter.

Chapter 8. General Discussion

The focus of this thesis was to investigate the effect of test preparation on educational test performance as well as Chinese students' strategy use in test preparation contexts. The fundamental aim of the research programme is to develop a general understanding of students' study during the test preparation process and its effect. This research synthesises theories regarding high-stakes testing and test preparation in the educational assessment field, as well as theories regarding students' learning in the fields of cognitive and metacognitive science, to provide support for the research design and data interpretation. This research includes several interrelated studies, the results of which are mainly presented in three individual research papers (Part II). Some additional and cross-study analyses were outlined in Chapter 7.

This chapter first extracts key evidence from findings to answer the overarching research questions and discusses how findings in this research parallel or differ from prior studies. Next, it revisits the purposes of the study and addresses possible implications of this research. Then it discusses the limitations of the current research project and possible directions of future research. This thesis finishes with concluding remarks.

8.1 Discussion of the overarching research questions

This section consists of four sub-sections to discuss findings around the three overarching research questions. Note that the first research question has been addressed in the first two subsections, synthetical and empirical findings are discussed separately.

- RQ1. What effect do test preparation interventions have on test performance?
- RQ2. What factors affect the effectiveness of test preparation interventions?
- RQ3. How do Chinese students regulate their use of testing and restudy? What factors influence their strategy choices?

8.1.1 What effect do test preparation interventions have on test performance?

The first research question of the current research was, 'What effect do test preparation interventions have on test performance?'. The literature reported various results on this question. To gain a synthetic understanding, a meta-analysis has been conducted in this research to quantify the average effect of test preparation on large-scale educational tests based on previous experimental studies. In general, a significant and positive test preparation effect was found ($g = 0.38$, $p < .001$) on criterion tests. According to Cohen's rule of thumb for describing the magnitude of effect size, this effect size was small to medium (J. Cohen, 1988), but it is very close to the hinge point (0.4) that is typically considered as 'practically' significant in educational interventions (Hattie, 2008).

This effect size appears to be larger than effects reported in two previous meta-analyses on test preparation effects. Bangert-Drowns et al. (1983) reported an effect size of 0.25 for coaching programmes on achievement test performance, while Kulik et al. (1984) reported an effect size of 0.15 for coaching on the SAT. The most direct reason why the effect reported by the current study differs from previous studies is that the main body of included cases in the current meta-analysis is different. The two previous meta-analyses focused on intelligence tests and college admission tests in the U.S., but many of the included articles of previous meta-analyses were excluded by the current study, because some of them did not use large-scale educational tests as the measure (e.g., Boger, 1952) or did not provide enough details about intervention procedures (e.g., Bajtelsmit, 1977), or because they preceded the date range specified by the inclusion criteria. Any SAT-related studies conducted before 1978 and any ACT-related studies conducted before 1985 were excluded, as example items of the two tests were not provided by the testing bodies to the public prior to the two years (*A (Mostly) Brief History Of The SAT And ACT Tests*, n.d.), so teachers and students had very limited access to test-related materials. The two previous meta-analyses (Bangert-Drowns et al., 1983; Kulik, Bangert-Drowns, et al., 1984), however, were conducted in 1980s and their

included articles were mostly published between 1960s and1980s. In contrast, the current meta-analysis includes more recent studies, where teachers and students may have used a vast array of test preparation materials, many of which are free and publicly available online. Using the well-designed test preparation handbooks and the sample tests published by test developers, teachers and students may tailor their instruction and learning to the tests, leading to greater score gains.

Note that about one third of the included studies in the current meta-analysis examined the test preparation effect for large-scale English proficiency tests, such as IELTS and TOEIC, and they reported a relatively larger effect size of 0.47 (although not significantly greater than preparation interventions on college admission tests and other tests). This larger effect size on English proficiency tests may also be explained by teachers and students' familiarity with the format and tested content on criterion educational tests. As there is a huge market for preparation for English proficiency tests, especially for those tests widely used by international students for purposes of admission to higher education, relevant test preparation products are flooding the market, so that teachers and students can make full use of relevant materials to improve their familiarity of tested content and test format. Besides, due to the high frequency of some language tests (e.g., there can be up to four or five IELTS and TOEFL tests in a month), their predictability might be higher than for other infrequently held tests. For instance, Yu et al. (2017) found that in China, TOEFL teachers from test preparation schools routinely rotated to take the test and recorded test items. Their test-taking experience and those shared by other test-takers are known as 'Ji-jing' (机经, test-taking experience). Test preparation schools frequently collected and analysed Ji-jing to make predictions about future test items and provide students with model essays or sentence templates to memorise. While this approach has great potential to improve test performance,

it may undermine the validity interpretations of test scores, as the inflation of test scores cannot reflect students' improvement in English language proficiency.

Therefore, we should be careful about the interpretation and generalisation of the test preparation effect derived from the meta-analysis in Paper 1. It is still questionable whether the preparation for a specific test can produce a beneficial impact on students' authentic learning and whether the effect can be transferred to other tests. Two included studies (e.g., Hardison & Sackett, 2008; Trenkic & Hu, 2021) used other tests or tasks with similar domain or general proficiency tests as additional measures, and their results suggested that the effect of preparing for a specific test was not transferred to other tests. Based on the Transfer-Appropriate Processing (TAP) framework, it is possible that the content and processing strategies during test preparation cannot match those required by the tests. Thus, the transfer is far and hard to achieve. However, due to insufficient data, this finding is not conclusive. More studies are needed to verify the transferability of test preparation effects, so as to help us better interpret the underlying mechanism of the test preparation effect, i.e., to what extent the score gains can be attributed to the acquisition of broad knowledge or the improvement of general cognitive skills.

As mentioned above, the meta-analysis in Paper 1 concerns the test preparation effect on large-scale educational tests and most included studies demonstrated the effect of teacher-led test preparation activities (i.e., coaching). Thus, results of this study only give a partial understanding of the effectiveness of test preparation. The effect of student-driven test preparation activities, and the test preparation effect on classroom tests, are still unclear. The use of classroom tests is more widespread in educational practices. It is widely acknowledged that they are useful formative tools and can benefit students' learning (Black & Wiliam, 1998). Classroom tests can also be used for summative purposes. For example, teacher-made tests are widely used in higher education, as students need to pass those to earn a specific number of course credits. However, there is relatively limited evidence of the magnitude of

the test preparation effect on these tests, probably because the difference between test preparation and students' learning is blurred and most previous classroom studies did not clearly distinguish between them.

Students' everyday learning behaviour can be different from their learning when tests are approaching, as observed in many previous studies (e.g., Blasiman et al., 2017). Although researchers have identified that spaced learning is superior to massed learning where students learn a large amount of information within a short period of time (Kornell, 2009; Smolen et al., 2016), in students' practice, they still tend to cram before the test, and mainly rely on restudying their notes and texts to prepare for exams (Blasiman et al., 2017). The four (quasi-) experiments in this research shed light on the effectiveness of 'massed' test preparation before the final test, as well as probe which strategy is more effective in the 'massed' test preparation session. Using a continuously cumulating meta-analysis to synthesise the effects of the using different test preparation strategies across the four experiments, the results indicated that the massed test preparation practices had a significantly positive effect on post-tests which were taken immediately after the interventions ($g = 0.60$, $p < .01$), but the effect was not significant on delayed course test. This finding is in line with previous literature on the testing effect (e.g., Kornell, 2009), and it can be inferred that the effect of massed practice cannot benefit long-term learning. What practices can better enhance long-term learning remains a question that requires further exploration. The complexity of real-world education practices and various factors related to learning makes it difficult to establish causal relationships between specific practices and long-term effects. An analysis of 141 large-scale educational Randomised Controlled Trials (RCTs) commissioned by large educational funding bodies in both UK and the U.S. found that interventions aimed at improving academic achievement in K-12 contexts had a very small effect size (0.06 SDs) on average. This makes the effect of test preparation all the more impressive.

8.1.2 What effect do test preparation interventions have on test performance?

Regarding the effect of specific strategies in the 'massed' test preparation session, three (quasi-) experiments in the research compare the effects of two most used test preparation strategies – testing and restudy– on post-tests scores and the delayed course test performance, as well as whether the effects differ in different test conditions (high-stakes condition *vs.* low-stakes condition), and whether students' monitoring and control of their learning amplify the effects. It has been found that testing practice with feedback is an effective strategy in high-stakes test preparation conditions, yielding large effects on both the post-test and on the delayed course tests when compared to the control condition. In contrast, testing practice without feedback failed to lead to significantly better performance in both the post-test and the course test.

The findings suggest that feedback plays a significant role in the process of test-enhanced learning. This conclusion reinforces the established body of research that has consistently emphasised that feedback benefits retention and long-term learning, and can facilitate the transfer of learning (McDaniel et al., 2007; Pan & Rickard, 2018; Vojdanoska et al., 2010). The three experiments were conducted in a real test preparation context and used materials from the real curriculum and previous tests. No identical item was used across the learning and testing stages of the interventions, and such a design required a certain degree of transfer. It is believed that feedback provided to students after the retrieval practice can help students identify their learning status, as well as the underlying concepts and principles that are assessed by the task. Therefore, students can better encode the testing item and answer, and are more likely to be able to apply their knowledge of this task to other contexts (Pan & Rickard, 2018).

Restudy was also an effective strategy, which produced significant effects on post-test in the high-stakes condition. However, the restudy effect does not seem to facilitate long-term learning, as evidenced by the finding that the use of restudy did yield significant score

improvement on the delayed course test compared to the control condition. It should be noted that the findings from the third experiment reveal that students' monitoring and control of their restudy practice might improve the potential of restudy practice on long-term learning. Metacognition and self-regulated learning theories suggest that ideally a student who recognises knowledge gaps or difficulties in understanding will invest more time and effort (Winne & Hadwin, 1998). Students in metacognition groups in Experiment 3 showed more restudy attempts, which might reflect students' regulation of their learning. However, findings also suggest that students' monitoring accuracy did not improve, and the metacognition groups did not achieve greater gains after the intervention, thus this research does not provide enough support for the hypothesis regarding metacognition.

When comparing the effects of testing practice and restudy, no statistical difference was observed in both immediate post-tests and the longer delayed course tests. This result looks surprising at the first glance, as it is inconsistent with a large body of literature confirming the superior effects of testing over restudy in classroom settings (e.g., Agarwal, 2019; Roediger et al., 2011). However, as discussed in Paper 2, these studies mainly used identical or highly similar materials in the intervention and testing. In comparison, no identical item was used between the intervention stage and the testing stage in the current research, and the items in criterion tests assessed not only knowledge recall, but also knowledge understanding and application. This might explain why testing did not show advantages than restudy. Similar findings have been reported in problem-solving tasks in mathematics or science, where the benefits of worked examples (restudy) have similar or even better effect than problem-solving practices (testing), (van Gog et al., 2011; van Gog & Kester, 2012; Yeo & Fazio, 2019)

The knowledge-learning-instruction framework proposed by Koedinger et al. (2012) provides a valuable perspective to understand effects of the two strategies in educational contexts. As suggested by the framework, the fit between the nature of knowledge

components and the learning processes required to acquire the knowledge is critical for learning outcomes. Based on this, the effects of testing and restudy practices are determined by what kinds of knowledge are being acquired. For factual knowledge (constant-constant knowledge components), the testing practice can be particularly effective to consolidate the long-term retention, while restudy or passive re-reading the information is less effective. When it comes to more complex materials, for example, when students are required to acquire knowledge that may have variable application conditions or cues and variable responses (variable-variable knowledge components), such as problem-solving tasks, restudy can be as effective as testing, or even more effective. This is because complex knowledge components necessitate predominantly induction or compilation processes and restudy allows the learner to build up a more comprehensive understanding of the procedure. In contrast, the testing practice may not provide as much opportunity for learners to integrate information into a deeper understanding as it focuses on recalling specific information.

8.1.3 What factors affect the effectiveness of test preparation interventions?

To better understand the test preparation effect, this research examines the factors that can influence the magnitude of test preparation effect. Synthesising findings across studies, two types of factors have been identified, which could moderate the test preparation effect.

Test preparation characteristics

The meta-analysis suggests that the narrower the test preparation intervention, the greater the effect on test performance. Findings from (quasi-) experiments also provide some support for this notion, as the interventions focused similar though not identical content with the test and yielded a relatively larger effect size on immediate tests. These findings indicate that students' increased domain-specific knowledge can lead to better test performance (Arendasy et al., 2016). Besides, the transfer-appropriate framework and cognitive load theory also provide helpful insights to explain the benefits of narrowed test preparation

intervention on test performance (Pass et al., 2003; Sweller, 1994, 2010; Morris et al., 1977). Test preparation instructions include broad knowledge or some general cognitive skills which cannot be directly applied to students' test-taking might cause ineffective, extraneous cognitive load and thus hamper learning. However, effective cognitive load that focuses on activities relevant to the acquisition of required knowledge or skills is germane to learning. Therefore, in order to be effective on test performance, especially for authentic educational assessment tasks that impose a high intrinsic cognitive load, test preparation instructions and practice should be designed to minimise the extraneous cognitive load. Using materials that directly tapped the test-related content could allow learners to use the freed-up cognitive capacity to process knowledge and skills. Additionally, the transfer-appropriate framework states that when the processing demands of the learning task match the processing demands of the transfer task, transfer will be more successful. Therefore, for interventions with a narrow focus, for which practice materials provided during the intervention and the testing materials in the test stage are largely matched, enables students to process the material in a transfer-appropriate manner.

The use of workbooks, revealed by the meta-analysis, is another significant moderator of test preparation effects. As discussed in Paper 1, a typical workbook mainly introduces substantial information about the test, including the test structure, test format, question types and sample test questions, as well as providing some practice tests and useful test-taking strategies. Some workbooks also help students to review frequently tested knowledge in a structured and organised way. Therefore, using workbooks can improve students' familiarity of the test, as well as help solidify students' understanding of the test-related material and improve their retention of the information.

The meta-analysis of test preparation effect also suggests that instructing students on how to manage test preparation process through managing time, goal setting, analysing rubrics and evaluating one's responses, etc., is a significant moderator. This finding

highlights the potential contributions of utilising metacognitive strategies in the test preparation process. Although the (quasi-) experimental studies did not confirm the additional effects of using specific metacognitive strategies of monitoring and control on the test preparation effect, this does not mean that metacognitive strategies are ineffective. Findings of Experiment 3 already indicated that using metacognitive strategies plus restudy practices has the potential to improve the restudy effect. Besides, given that both the instruction and practice on metacognition was short, students might not be well trained, and their metacognitive accuracy might not be improved. Future studies are needed to explore whether the improvement of metacognitive accuracy after longer (or multiple) interventions bolster the effectiveness of the use of other learning strategies.

Consistent with previous studies (Arendasy et al., 2016), it is found that teaching students test-taking skills is another significant moderator of the test preparation effect. Utilising skills such as guessing, eliminating incorrect answers, organizing the responses in clear ways can increase the possibility of getting more scores, although these skills cannot influence the underlying construct highlighted by the test.

Additionally, the continuously cumulating meta-analysis (CCMA) on the four experiments revealed that the interval between the test preparation session and the test was the only significant moderator of the effect across the four studies: the closer the test, the better the effects of test preparation. This might be because memory retention declines over time. In actual test preparation contexts, cramming the night before the exam is a common approach (Appelrouth et al., 2017; Blasiman et al., 2017). Although this finding provides support for the effect of cramming on immediate test performance, it does not mean that cramming is an effective long-term learning strategy. Cramming a large amount of information in a short time may overload the working memory and reduce the ability to transfer information to long-term memory. Moreover, as indicated by the findings, long-term learning is difficult to be improved through cramming practice. Spacing practice, is a more

effective for memory retention and long-term learning, as confirmed by many researchers (e.g., Kornell, 2009; Rawson et al., 2013).

Individual characteristics

Intriguingly, the meta-analysis showed that students who had **previous test-taking or test preparation experience** can experience more benefits of test preparation than those who had not. This is possibly because students with some exposure to the test do not need to spend their time and cognitive load on test familarisation. This speculation has been built in the literature on the retest effect. For example, Anastasi (1981) hypothesised that test takers who are not familiar with the test need to familiarise themselves with the test situation while solving the first few test items, and this make them perform a dual task while solving the first few test items. Thus, less-familiar test takers show poorer test performance in the first few tasks than more test-familiar test takers. Anastasi (1981) suggested that practice testing or solving sample items can reduce the detrimental effect caused by unfamiliarity. However, this hypothesis received less evidence from the literature on the retest effect. Considering that test preparation is a more complicated process that involves more activities than simply retesting, our findings suggest that a certain degree of familiarity of test-taking and test preparation can save students' cognitive resources and enable them to identify what knowledge is highlighted by the test and how the knowledge is assessed, thus they are better able to gain a robust grasp of the knowledge and some experience in problem-solving, which further helps them to apply what they have learned to solve different test questions in a new testing context. This potential explanation requires more research before it can be concluded firmly.

Participants in the four (quasi-) experiments in the current research were from the same school and took the educational psychology course provided by the same teacher,

thus their experience of test-taking and test preparation can be regarded as similar. However, it is assumed that **students' prior attainment level** might have moderated test preparation effect, as indicated in some previous findings (Carpenter et al., 2016).

The relationship between students' prior attainment levels and their score gains have been explored in the additional analyses in this thesis. A general pattern has been identified across the four (quasi-) experiments: low performers achieved significantly better score gains in the immediate post-test than high performers. This finding can be explained in various ways. First, the existence of ceiling effect has been considered, as high-performing students already reached a certain level of mastery and thus their learning progress may reach ceiling. However, it is noticed that in the four (quasi-) experiments, most high performers only scored between 50% and 60% of the total score in the pre-test, and their scores in the post-test and course test reached about 80% of the total score. Such an increase seems to be at odds with the assumption of a ceiling effect.

A second possible explanation is the occurrence of the expertise reversal effect, which refers to the reversal of the effectiveness of instructional techniques on learners with differing levels of prior knowledge (Kalyuga, 2009). Based on cognitive load theory, the effectiveness of teaching and learning depend on the characteristics of learners' cognitive process that vary with the changes in their domain-specific knowledge base. For instance, when a student with prior knowledge and relevant experience works on a familiar type of test question, their relevant schemas and routine problem-solving procedures are activated to solve the question and intrinsic load are generated. If the student continues to restudy the exemplar answer or worked examples to the question, external support materials then cross-reference with already-learned procedures. The student could therefore be distracted from fluent cognitive processing that takes full advantage of their existing knowledge structures and might have working memory overloaded by additional, extraneous activities. In contrast, the study could be

more effective for novices without enough prior knowledge and experience (Renkl, 2005). In this research, when high-performing students reviewed relevant topics, they may try to cross-connect what they already knew with the provided information, while this process, however, increased the extraneous cognitive load. In terms of low-performing students, the interventions with clear definitions of educational psychology concepts may generate productive intrinsic load, which was essential for constructing missing schemas in their knowledge structure. Therefore, students with lower prior attainment benefited more from the interventions.

Another possible reason for why students with lower prior attainment level benefited more could be attributed to the nature of the interventions. The interventions in the four (quasi-) experiments in this research focused on factual knowledge of educational psychology concepts. Students with higher prior attainment had possessed greater retention on relevant factual knowledge, thus they benefited less from repeatedly studying. In comparison, lower prior attainment who lacked a fundamental understanding of the subject matter could benefit more through learning factual knowledge during the interventions. Future research can further explore this explanation by providing different types of knowledge (e.g., procedural knowledge or application tasks) during the intervention to students with different ability levels and compare the effects.

8.1.4 How do Chinese students regulate their use of testing and restudy? What factors influence their strategy choices?

In the first three experiments, the test preparation strategy was manipulated, and students were required to use one specific strategy to prepare for the test based on the group allocation. Although students were allowed to learn the material multiple times, and in Experiment 3 students were even reminded that they could revisit each topic, which somewhat approximated to real learning situations, the manipulation of strategy use could not reflect students' actual test preparation practices. Therefore, in Experiment 4, students were

allowed to use testing and restudying interchangeably and revisit the material using either strategy as many times as they needed.

Through analysing how students approached each to-be-learned topic during the test preparation in Experiment 4, possible patterns of students' study have been identified. Overall, students used one-time testing to review a topic most frequently (about 34%), followed by using both restudying and testing in turn (about 30%), while students rarely used repeated restudying (less than 4%).

Students' resistance to repeated restudying has also been confirmed in the third experiment in Paper 2, as demonstrated by the fact that students in restudying groups made fewer study attempts. Moreover, given that students were allowed to make multiple study attempts on each topic, the descriptive analysis showed that students chose testing and restudying equally frequently in their initial study attempts when they reviewed educational psychology topics. While in their second and third study attempts, students chose testing more frequently rather than restudying. These findings strongly suggest that students in this research preferred the testing strategy. Qualitative data collected following the experiments further provide further support for this conclusion, illustrated by the fact that the majority of participants directly expressed their preference for testing.

Intriguingly, students survey responses suggested that while many students did find restudying to be a useful learning tool to facilitate the development of a complete knowledge structure, they believed that the efficacy of restudying needed to be linked with testing practices to be fully achieved. Specifically, some students stated that restudying the knowledge that cannot be answered correctly in prior testing practices was an optimal way; while some other students mentioned that they would first restudy the poorer learned materials then test themselves to further consolidate their understanding of knowledge. These perceptions reveal that Chinese students recognise the role of restudy as a learning tool, and

they recognise two possible functions of testing, that is testing as a monitoring strategy and as a learning strategy.

Previous studies have consistently concluded that students use testing as a monitoring tool rather than a learning tool (Geller et al., 2017; Hartwig & Dunlosky, 2012; Karpicke, 2009; Kornell & Bjork, 2007; Rivers, 2021), as evidenced by the widely noticed pattern that students tend to choose testing practice when successful retrieval is more likely to occur (for a review, see Rivers, 2021). For example, some experimental studies have confirmed that students chose to restudy more often for difficult items or when they were less confident, but as the difficulty level decreased or as their confidence increased, testing was chosen more than restudying (Badali et al., 2022; Toppino et al., 2018; Tullis et al., 2018). The current research yielded results that closely align with those of previous experiments. Since 43% of first study attempts were followed by a second try, data on students' strategy choices in the two attempts were substantial, which enabled the analysis of potential factors that could influence students' strategy choices. This research suggests that students' confidence in the topic and how the topic would be tested (i.e., by multiple-choice questions or by short response questions) predicted which strategy students chose. Students tended to test themselves when they were more confident in their knowledge mastery of the topic, while restudying the texts of the topic when they were less confident. Additionally, students tended to do multiple-choice questions rather than short-response questions. When they knew the topic would be tested by short response questions, they were more inclined to restudy the texts. Moreover, the strategy choice in the first study attempt was also a predictor of students' strategy choices in their second attempts (if they engaged in a second one). Students tended to choose a different strategy in their second attempts than their first one. To some extent, all these findings support the inference that students in the current research also use testing as a monitoring tool.

However, there is also ample evidence that in this research that testing was not only used for monitoring. If participants in this research only employed the testing strategy to diagnose the gaps in their knowledge (i.e., answering the test questions incorrectly), they would have used restudying to learn the poorly acquired knowledge after unsuccessful retrieval attempts. Equivalently, students would have stopped studying items after one correct recall, as evidenced by some findings in several studies (Badali et al., 2022; Rawson et al., 2013). However, the current research found contrary results. When students chose to do testing as a first study attempt and wrongly answered the testing question, in 46% of the cases they dropped the topic and in 30% of the cases they re-did the test question (39%). Besides, when students had answered the testing question correctly, there were still about 14% of cases where students revisited the topic using either restudying or testing. These behaviours hint at a new speculation: Chinese students use testing as not only a monitoring tool, but also as a learning tool.

Based on similar results identified in their study, Toppino et al. (2018) claimed that the potential benefits produced by testing might be somewhat understood by students, but they may also think that successful retrieval is necessary to reap those benefits. However, this assumption is not substantiated by the current research. It seems that unsuccessful retrieval has also been highly valued by students. Qualitative data suggested that some students believed that mistakes and feedback were important learning resources, because with feedback provided, a failed retrieval could leave a deeper impression and even lead to better retention after encoding the corrective feedback.

To summarise, in the current research, students showed an obvious preference for the testing strategy and a strong trust in its effect. This is at great variance with previous findings based on student surveys (Blasiman et al., 2017; Karpicke et al., 2009; Kornell & Bjork, 2007; Kuhbandner & Emmerdinger, 2019), which indicated that rereading was largely preferred over testing and students underestimated the benefits of testing. It is plausible that

this variation arises from the different ways that data were collected, as it is long claimed that self-reported behaviours do not always match up with the actual practice (Dang et al., 2020). However, as discussed in Paper 3, compared to Jane et al.'s (2018) study, which also gathered data from experiments and where participants used a relative low frequency of testing, participants in the current research still often selected testing. Therefore, students' preference for the use of testing seems to be related to the social-cultural context where the study was conducted. Previous studies mentioned above were mainly conducted in the U.S., where the education system and educational practice are distinct from those in China. As mentioned in previous chapters, the reliance on testing practice in the Chinese educational system and the social awareness of the benefits of testing (e.g., the nationwide discussion around the 'small-town swot') are likely to inspire students' trust in the value of testing.

Another possibility for students' frequent use of testing was that students prioritised the efficiency of test preparation practice, thus they would choose strategies that are regarded as more time efficient. As indicated by the survey, students tended to believe that using testing can help them to cover as much learning material within a limited time. Besides, answering testing questions could help them refresh the potential tested content and improve their familiarity with how to construct responses to test questions.

However, as discussed in previous sections in this chapter, students blended use of the two strategies did not yield greater score gains than the use of one particular strategy (whether or not metacognitive training and tasks were provided). This result is not surprising given how heavily students relied on the testing strategy discussed above. As revealed by both the meta-analysis and previous experiments, the use of testing was not a significant moderator of test preparation effect. Thus, increasing the amount of testing practices cannot necessarily magnify educational test performance. Contrarily, restudy is an equally effective strategy with testing, and Experiment 3 also sheds light on the potential arising from better regulation of restudy practices. Taking these findings together, it can be inferred that Chinese

students overestimated the effect of testing practice and that their excessive reliance on testing could not produce the desired benefits.

Another possible reason why the combined use of the two strategies did not yield more benefits was that students did not regulate their study effectively. Actually, some traces of students' dynamic regulation of their study have been observed in both Experiment 3 and 4. For example, it has been found that students' modified confidence judgements after their initial study attempt, rather than their initial judgements, predicted their second strategy choice, and this can be regarded a dynamic monitoring and control process. However, the observed regulated study behaviours do not necessarily mean effective self-regulated learning, as evidenced by the finding that students' monitoring accuracy was not improved in Experiment 3 and 4.

Previous studies stated that study strategies endorsed by students are related to academic performance: high-performing students tend to choose more effective strategies (Hartwig & Dunlosky, 2012). Geller et al. (2017) found that high-performing students (relative to low-performing students) were more likely to endorse self-testing than rereading. The current research found an opposite pattern: students with a low prior attainment level tended to choose rereading more frequently than those with a high prior attainment level. However, neither the frequency of the adoption of a particular strategy nor the overall study attempts during the intervention led to better test performance. The reason why low-performing students tended to use restudy to go over the material and to familiarise themselves with the content rather than use testing may be that using testing on poorly learned knowledge would be cognitively demanding and perceived as less effective.

8.2 Implications

Possible implications of this research programme for both research and practice are presented in this section.

8.2.1 Implications for research

This integrated research provides both synthetical and empirical evidence to understand test preparation effect in general and the effect of students' self-use of two specific test preparation strategies in particular. Notably, this study also reveals some distinctiveness in how Chinese students use and perceive the testing strategy in their practice, which further calls for more future study to examine students' study behaviours in different socio-cultural contexts, and if possible, to conduct cross-cultural comparisons.

Overall, this research contributes to knowledge in four ways: First, by conducting a meta-analysis, the study provides a comprehensive understanding of the overall estimate of effects of test preparation in large-scale educational tests, as well as whether factors of test characteristics, test preparation practices, and test takers' characteristics moderate the direction and magnitude of test preparation effect. Although there are some precedents of using meta-analytic methods to evaluate test preparation effects (Bangert-Drowns et al., 1983; Kulik, Kulik, et al., 1984), they were published decades ago and were mainly conducted in the U.S., focusing solely on the effect of a single practice (i.e., coaching or testing) on performance of educational tests or intelligence tests. The current meta-analysis is more inclusive, taking a wide range of test preparation activities into consideration and including relevant studies conducted in different regions around the world into the analysis, thus reporting more comprehensive and up-to-date findings that can be generalised to wider population and various educational contexts.

Besides, the current research can make a contribution to existing literature and real-world test preparation practice by using a rigorous experimental design to investigate the

effectiveness of test preparation. A large number of previous experimental studies lacked control and rigor, since most studies evaluated the effect of long-term test preparation courses while they reported little details of teachers' and students' test preparation practices during courses, and some studies did not involve a control condition. The results might be inflated and unconvincing, and it is difficult to differentiate the contribution of a specific strategy or material usage to the overall test preparation effect. In comparison, experiments in this research employ a rigorous design and manipulate test preparation conditions by examining the use of two commonly used strategies—testing and restudying, with other factors being controlled. Therefore, this study yields more reliable findings and can provide meaningful insights to real-world test preparation practice.

In addition, the findings of this research have great potential to contribute to the literature on testing effects. Most studies on the testing effect were conducted in laboratory settings, using word pairs or reading pieces as materials to examine recall and retention. Testing effects are less investigated in classroom settings, especially in real test preparation settings using educational assessment tools. The experiments in this research are conducted during a test preparation session of an educational psychology course in a teacher training college in China, providing empirical findings of testing effect in an ecologically valid context and explore whether testing can be beneficial to not only knowledge recall but also knowledge transfer. Therefore, the findings of this research can be a substantial supplement to current literature on testing effects and will expand the scope of research in this area.

Last but not least, this research could also contribute to the literature by taking students' regulation of strategy use into account. Students are the most important participants in test preparation, but their self-initiated practices have not been deeply investigated. By analysing how students monitor their learning status and control their choices of strategy use in the test preparation context, we could better understand how students perceive the functions and effectiveness of different test preparation strategies and their potential to regulate their own

test preparation process. Moreover, the findings indicated that unlike their Western peers, Chinese students recognise the role of testing as both a direct learning method and as a monitoring tool. This finding differs from previous research and illustrates the importance of understanding students' preferences and perceptions of study strategies in a sociocultural context.

8.2.2 Implications for practice

The thesis provides an overall review of the effect of test preparation and Chinese students' strategy use during the process. Findings from the meta-analysis and (quasi-) experiments reveal some possible factors that can contribute to greater score gains from test preparation practice, which are important implications for prospective teachers and students who are or will be getting involved in test preparation practices.

The meta-analysis of the effect of test preparation on large scale educational tests presented in Paper 1 confirmed the positive moderating effects of using workbooks, learning test-taking skills, and learning test preparation management skills. These materials and strategies should be given consideration when teachers design test preparation courses and when students plan their self-study. Besides, though cramming practice focused specifically on the test-related content can improve test scores, its long-term effect on test performance can be severely compromised, and such an effect cannot be transferred to general context. Both teachers and students should fully acknowledge that the inflation of scores produced by cramming might not represent authentic improvement of learning, and thus avoid the reliance of intensive test preparation sessions in real teaching and learning practices.

When designing test preparatory courses, teachers should take students' prior experience and their prior attainment level into account. It may be more beneficial to provide low-performing students with factual knowledge to establish a fundamental understanding of knowledge, while high-performing students may benefit more from practising tasks that required high cognitive demand. Additionally, it is important for novice test-takers to know

they might not gain substantial gains from intensive test preparation courses right before the test.

The current research also provides potential suggestions for students' self-study for upcoming course tests. As indicated by findings, students can optimise their test performance utilising both testing and restudying practices right before the test. When testing practices are employed, it is important for students to use the corrective feedback to monitor their response and encode the feedback elaborately for a better mastery. When students reread a topic, it could be beneficial using some metacognitive strategies to improve the metacognitive awareness of their restudy, i.e., whether a full understanding has been built through the rereading, to improve the efficacy of the rereading practice.

The findings from (quasi-) experimental studies suggest that in terms of effects on authentic educational tests which assess not only knowledge recall but also understanding and application, testing practice is not superior over restudy. However, it seems that Chinese students might overestimate the benefits of testing as they have shown apparent preference on the use of testing in their study. This thesis implies that Chinese students may need to re-examine the effectiveness of testing more sensitively and optimise their leaning by taking advantage of multiple strategies.

8.3 Limitations and Future directions

There are five main limitations in the current thesis that require attention.

The first limitation is about the generalisability of the results from both the meta-analysis and (quasi) experiments. In terms of findings in the meta-analysis, as most included studies were carried out in the U.S. context and focused on college admission tests, though findings indicate that the magnitude of the effects in the U.S. was similar in other contexts, and the difference between different tests was not significant, some caution should be taken when generalising results to other testing contexts, especially tests of other domains (i.e.,

science, medicine, foreign language except English) or in less investigated regions (i.e., regions in the Middle East, Southeast Asia, and Africa). Regarding the four (quasi-) experimental studies, as they were conducted in a teacher training college in southwest China, where some student enrolled in this school with higher *Zhongkao* scores because of the Free Teacher Education (FTE) policy, while other self-funded students had relatively lower *Zhongkao* score and could not get admitted to key high schools. Therefore, there was a certain level of variation of students' ability levels in this research which guaranteed a degree of representativeness of the population. However, the motivation of students in such schools may be different from that of students in regular high-schools or universities as can be inferred from the averagely low test scores in pre- and post-tests and the course tests, i.e., the average scores were less than 60% of the overall score in most cases. Additionally, as participants in this study were enrolled in educational psychology courses, they acquired knowledge about learning strategies through the lectures of the courses. Thus, they were more likely to be familiar with the effectiveness of testing practice and metacognitive strategies. Though the averagely low pre-test scores indicated that participants in this research programme may not have applied course content to their own strategies, it remains a possibility that they may tend to choose testing practices more often in their self-study, and the metacognitive training effect may have weakened due to their existing level of metacognitive awareness of their learning status. It is important to note that similar studies conducted with participants from other subject domains may yield different results. When generalising the findings of this research to the wider population of Chinese students, it is necessary to take into account the context in which this research was conducted. Future research which includes a more representative sample from regular schools or universities in academic tracks as well as participants from different subjects can further justify findings of this research and add to a broader understanding of test preparation effect in China's contexts as well as Chinese students' behaviours and perceptions.

Another limitation pertains to the interpretation of the course test results in the (quasi-) experimental studies. These studies were conducted in an authentic educational setting, students in Experiments 1, Experiment 3, and Experiment 4 prepared for real high-stakes tests when they participated in the study. To make sure that this experiment would not disadvantage any students from particular groups, links to the online materials used during interventions were shared with all participants after the experiments. Although students' access to the online materials were recorded, the majority of students did not provide information about which test preparation group they were from in the experiments, thus it was unable to identify the extent to which relevant materials were reviewed. In addition, some students might use other materials (e.g., notes, the textbook) to prepare for the test on their own, and the information on their exposure was not collected. Therefore, students' performance on course tests cannot be fully attributed to the interventions. The effect of the intervention on the final course test results may be compromised, and students' performance is likely to have been affected by how well they prepared on their own after the intervention. Future studies can conduct ongoing follow-up records, examining whether students' strategy use change after the intervention, so as to understand the impact of training the use of specific strategies on students' understanding and usage of particular strategies, as well as the subsequent impact on students' achievement.

Additionally, in this research, restudy has been regarded as an integral strategy. However, students might use varying techniques within the restudy conditions, such as re-reading texts, taking notes, rehearsal, etc. (Kuhbandner & Emmerdinger, 2019). Therefore, the internal validity of the restudy condition could be improved in future research by addressing these different approaches and differentiating their effects. As revealed by findings of the current research, restudy is also an effective test preparation strategy that can improve students' performance on immediate tests, while a good regulation of restudy behaviours seems to have potential on improving long-term retention. Therefore, it would be

valuable for future research to break down the restudy strategy and investigate the individual difference in the use of specific techniques in the restudy condition, to gain a further better understanding on whether individual differences are related to the effectiveness of restudying.

The fourth limitation concerns students' blended use of the two strategies. In previous research, very limited research has investigated the optimal combination of strategies. Experiment 4 in the current research project has recorded and analysed students' blended use of two strategies in a test preparation session, presenting the dynamic strategy choices, strategy use patterns, predictive factors, and their beliefs on the effectiveness of strategies. Though this study illustrates a fine-grained picture of the what, when and why of students' study decisions for test preparation, no clear relationship between these decisions and test performance has been identified. More systematic experiments are needed to further explore whether certain types of strategy combination are associated with better performance, and whether the benefits change as a function of test content, test type, test structure and so forth.

Very limited qualitative data were reported in this research, and this is the fifth limitation. In terms of students' use of strategies for test preparation, it is suspected that at least some students might adaptively adjust their study decisions to different tests and also to different course (as a function of different class content and curriculum structure). Students' different use of strategies under different contexts can be better understood by proper qualitative studies which conduct interviews, do classroom observations, or record students' learning diaries, which can add to the current research programme and fully explore the degree to which students do adapt their test preparation behaviours in an attempt to reach best test performance.

8.4 Concluding remarks

This integrated thesis focuses on the effect of test preparation in educational contexts as well as students' usage of strategies when they prepare for tests. One overall meta-analysis and

four (quasi-) experimental studies conducted in a teacher training school in China were included in the thesis. In general, test preparation has a significant positive effect on test performance on both large-scale educational tests and classroom tests, particularly those test preparation interventions focused on relevant content and skills or drilled on tested content. However, the effect of preparing for a specific test cannot be transferred to other tests assessing similar domain or general proficiency, and the effect is diminished on delayed tests. The use of both testing with feedback and restudy are effective test preparation strategies, and effect sizes produced from the two strategies are similar. Besides, the use of workbooks and the development of test-taking skills and time management skills have significant moderating effects of the effect size, as indicated by the meta-analysis.

This thesis also provides insights into Chinese students' beliefs of the functions of the testing and restudy strategies as well as their study decisions when they are required to coordinate the use the two strategies to prepare for the upcoming test. Overall, students showed an apparent preference on the use of testing strategy, while the strategy of rereading was rarely used repeatedly and was often accompanied by testing afterwards. Students' study decisions on strategy use were predicted by their judgement of learning and the way the knowledge was tested. It seems that Chinese students can recognise both the direct (testing as an effective learning tool) and indirect effect of testing (testing a monitoring tool to diagnose learning), which might relate to wide recognition of the benefits of testing practice in Chinese educational system and the ubiquitous use of testing in Chinese society.

References

A (Mostly) Brief History Of The SAT And ACT Tests. (n.d.). Retrieved 2 January 2023, from https://www.erikthered.com/tutor/sat-act-history.html#y1974

Abel, M., & Bäuml, K.-H. T. (2020). Would you like to learn more? Retrieval practice plus feedback can increase motivation to keep on studying. *Cognition*, *201*, 104316. https://doi.org/10.1016/j.cognition.2020.104316

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the Classroom: Teachers' Opinions of Statewide Testing Programs. *Theory Into Practice*, *42*(1), 18–29. https://doi.org/10.1207/s15430421tip4201_4

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

Alderman, D. L., & Powers, D. E. (1980). The Effects of Special Preparation on SAT-Verbal Scores. *American Educational Research Journal*, *17*(2), 239–251. https://doi.org/10.3102/00028312017002239

Alderson, J. C., & Wall, D. (1993). Does Washback Exist? *Applied Linguistics*, *14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement*, 35(1), 31–47. https://doi.org/10.1111/j.1745-3984.1998.tb00526.x

Allensworth, E., Correa, M., & Ponisciak, S. (2008). *From High School to the Future: ACT Preparation–Too Much, Too Late. Why ACT Scores Are Low in Chicago and What It Means for Schools*. CONSORTIUM ON CHICAGO SCHOOL RESEARCH.

Anastasi, A. (1981). Coaching, test sophistication, and developed abilities. *American Psychologist*, *36*, 1086–1093. https://doi.org/10.1037/0003-066X.36.10.1086

Anderson, F. T., & McDaniel, M. A. (2021). Restudying With the Quiz in Hand: When Correct-Answer Feedback is No Better Than Minimal Feedback. *Journal of Applied Research in Memory and Cognition*, *10*(2), 278–288. https://doi.org/10.1016/j.jarmac.2020.10.004

Andrews, S., Fullilove, J., & Wong, Y. (2002). Targeting washback—A case-study. *System*, *30*(2), 207–223. https://doi.org/10.1016/S0346-251X(02)00005-2

Anthenien, A. M., DeLozier, S. J., Neighbors, C., & Rhodes, M. G. (2018). College student normative misperceptions of peer study habit use. *Social Psychology of Education*, *21*(2), 303–322. https://doi.org/10.1007/s11218-017-9412-z

Appelrouth, J. I., Zabrucky, K. M., & Moore, D. (2017). Preparing students for college admissions tests. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 78–95. https://doi.org/10.1080/0969594X.2015.1075958

Arendasy, M. E., Sommer, M., Gutiérrez-Lobos, K., & Punter, J. F. (2016). Do individual differences in test preparation compromise the measurement fairness of admission tests? *Intelligence*, *55*, 44–56. https://doi.org/10.1016/j.intell.2016.01.004

Ariel, R., & Karpicke, J. D. (2018). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*(1), 43–56. https://doi.org/10.1037/xap0000133

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*(5), 258–267. https://doi.org/10.3102/0013189X07306523

Badali, S., Rawson, K. A., & Dunlosky, J. (2022). Do Students Effectively Regulate Their Use of Self-Testing as a Function of Item Difficulty? *Educational Psychology Review*, *34*(3), 1651–1677. https://doi.org/10.1007/s10648-022-09665-6

Baird, J.-A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: Fields apart? *Assessment in Education: Principles, Policy & Practice*, *24*(3), 317–350. https://doi.org/10.1080/0969594X.2017.1319337

Baird, J.-A., Hopfenbeck, T. N., Elwood, J., Caro, D., & Ahmed, A. (2014). Predictability in the Irish Leaving Certificate. *Oxford University Centre for Educational Assessment Report OUCEA/14/1*.

Bailey, D. R., & Judd, C. (2018). The effects of online collaborative writing and toeic writing test-preparation on l2 writing performance. *Journal of Asia TEFL*, *15*(2), 383–397.

Bajtelsmit, J. W. (1977). Test-Wiseness and Systematic Desensitization Programs for Increasing Adult Test-Taking Skills. *Journal of Educational Measurement*, *14*(4), 335–341.

Baker, D., & LeTendre, G. K. (2005). *National Differences, Global Similarities: World Culture and the Future of Schooling*. Stanford University Press.

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The Instructional Effect of Feedback in Test-Like Events. *Review of Educational Research*, *61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C. C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research*, *53*(4), 571–585. https://doi.org/10.2307/1170221

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication Bias: The Antagonist of Meta-Analytic Reviews and Effective Policymaking. *Educational Evaluation and Policy Analysis*, *34*(3). https://doi.org/10.3102/0162373712446144

Barber, L. K., Bagsby, P. G., Grawitch, M. J., & Buerck, J. P. (2011). Facilitating Self-Regulated Learning With Technology: Evidence for Student Motivation and Exam Improvement. *Teaching of Psychology*, *38*(4), 303–308. https://doi.org/10.1177/0098628311421337

Barnett, J. E. (2000). Self-regulated reading and test preparation among college students. *Journal of College Reading and Learning*, *31*(1), 42–53.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Batsell, W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological Validity of the Testing Effect: The Use of Daily Quizzes in Introductory Psychology. *Teaching of Psychology*, *44*(1), 18–23. https://doi.org/10.1177/0098628316677492

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*, 441–462. https://doi.org/10.1007/BF03173192

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further Synthesis and Appraisal. *Review of Educational Research*, *60*(3), 373–417. https://doi.org/10.3102/00346543060003373

Bishop, J. H. (1997). The Effect of National Standards and Curriculum-Based Exams on Achievement. *The American Economic Review*, *87*(2), 260–264.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-Regulated Learning: Beliefs, Techniques, and Illusions. *Annual Review of Psychology*, *64*(1), 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Black, P., & Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Blasiman, R. N., Dunlosky, J., & Rawson, K. A. (2017). The what, how much, and when of study strategies: Comparing intended versus actual study behaviour. *Memory*, *25*(6), 784–792. https://doi.org/10.1080/09658211.2016.1221974

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 657–668. https://doi.org/10.1037/0278-7393.15.4.657

Blazar, D., & Pollard, C. (2017). Does Test Preparation Mean Low-Quality Instruction? *Educational Researcher*, *46*(8), 420–433. https://doi.org/10.3102/0013189X17732753

Boger, J. H. (1952). An Experimental Study of the Effects of Perceptual Training on Group I.Q. Test Scores of Elementary Pupils in Rural Ungraded Schools. *The Journal of Educational Research*, *46*(1), 43–52. https://doi.org/10.1080/00220671.1952.10881998

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley and Sons. https://doi.org/10.1002/9780470743386

Bol, L., & Hacker, D. J. (2001). A comparison of the effects of practice tests and traditional review on performance and calibration. *Journal of Experimental Education*, *69*(2), 133–151. https://doi.org/10.1080/00220970109600653

Bookman, A. B. (1981). *The effects of preparation for tests on standardized mathematics achievement test growth* (Doctoral dissertation, University of Connecticut).

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously Cumulating Meta-Analysis and Replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. https://doi.org/10.1177/1745691614529796

Bray, M. (1999). *The Shadow education system: Private tutoring and its implications for planners*. UNESCO, International Institute for Educational Planning, Paris.

Bray, M. (2007). The Shadow Education System: Private Tutoring and its Implications for Planners, 2nd ed. UNESCO, International Institute for Educational Planning, Paris.

Bray, M. (2017). Schooling and Its Supplements: Changing Global Patterns and Implications for Comparative Education. *Comparative Education Review*, *61*(3), 469–491. https://doi.org/10.1086/692709

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS: 88. Chance, 14(1), 10–18.

Briggs, D. C. (2009). Preparation for College Admission Exams. 2009 NACAC Discussion Paper. In *National Association for College Admission Counseling*. National Association for College Admission Counseling. https://eric.ed.gov/?id=ED505529

Brunner, M., Artelt, C., Krauss, S., & Baumert, J. (2007). Coaching for the PISA test. *Learning and Instruction*, *17*(2), 111–122. https://doi.org/10.1016/j.learninstruc.2007.01.002

Broekkamp, H., & Van Hout-Wolters, B. H. A. M. (2007). Students' Adaptation of Study Strategies when Preparing for Classroom Tests. *Educational Psychology Review*, *19*(4), 401–428. https://doi.org/10.1007/s10648-006-9025-0

Buchmann, C., Condron, D. J., & Roscigno, V. J. (2010). Shadow Education, American Style: Test Preparation, the SAT and College Enrollment. *Social Forces*, *89*(2), 435–461. https://doi.org/10.1353/sof.2010.0105

Bunting, B. P., & Mooney, E. (2001). The Effects of Practice and Coaching on Test Results for Educational Selection at Eleven Years of Age. *Educational Psychology*, *21*(3), 243–253. https://doi.org/10.1080/01443410120065450

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604–616. https://doi.org/10.3758/MC.36.3.604

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, *13*(4), 273–281. https://doi.org/10.1037/1076-898X.13.4.273

Butler, D. L., & Winne, P. H. (1995). Feedback and Self-Regulated Learning: A Theoretical Synthesis. *Review of Educational Research*, *65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Byun, S. (2014). Shadow Education and Academic Success in Republic of Korea. In H. Park & K. Kim (Eds.), *Korean Education in Changing Economic and Demographic Contexts* (pp. 39–58). Springer. https://doi.org/10.1007/978-981-4451-27-7_3

Carpenter, S. K., Endres, T., & Hui, L. (2020). Students' Use of Retrieval in Self-Regulated Learning: Implications for Monitoring and Regulating Effortful Learning Experiences. *Educational Psychology Review*, *32*(4), 1029–1054. https://doi.org/10.1007/s10648-020-09562-w

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A Classroom Study on the Relationship Between Student Achievement and Retrieval-Enhanced Learning. *Educational Psychology Review*, *28*(2), 353–375. https://doi.org/10.1007/s10648-015-9311-9

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. https://doi.org/10.3758/BF03202713

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, *2*(2). https://doi.org/10.1177/2515245919847196

Cheng, L. (2005). *Changing Language Teaching Through Language Testing: A Washback Study*. Cambridge University Press.

Cho, E. Y.-N., & Chan, T. M. S. (2020). Children's wellbeing in a high-stakes testing environment: The case of Hong Kong. *Children and Youth Services Review*, *109*, 104694. https://doi.org/10.1016/j.childyouth.2019.104694

Cogliano, M., Kardash, C. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology*, 56, 117-129.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (6th ed). Routledge.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Crocker, L. (2003). Teaching for the Test: Validity, Fairness, and Moral Action. *Educational Measurement: Issues and Practice*, *22*(3), 5–11. https://doi.org/10.1111/j.1745-3992.2003.tb00132.x

Dahlke, J. A., Sackett, P. R., & Kuncel, N. R. (2023). Are Large Admissions Test Coaching Effects Widespread? A Longitudinal Analysis of Admissions Test Scores. *Applied Measurement in Education*, *36*(1), 1–13. https://doi.org/10.1080/08957347.2023.2172018

Damankesh, M., & Babaii, E. (2015). The washback effect of Iranian high school final examinations on students' test-taking and test-preparation strategies. *Studies in Educational Evaluation*, *45*, 62–69. https://doi.org/10.1016/j.stueduc.2015.03.009

Dang, J., King, K., & Inzlicht, M. (2020). Why Are Self-Report and Behavioral Measures Weakly Correlated? *Trends in Cognitive Sciences*, *24*. https://doi.org/10.1016/j.tics.2020.01.007

Davey, G., De Lian, C., & Higgins, L. (2007). The university entrance examination system in China. *Journal of Further and Higher Education*, 31(4), 385–396. https://doi.org/10.1080/03098770701625761

Devine-Eller, A. (2012). Timing Matters: Test Preparation, Race, and Grade Level1. *Sociological Forum*, *27*(2), 458–480. https://doi.org/10.1111/j.1573-7861.2012.01326.x

Domingue, B., & Briggs, D. (2009). *Using Linear Regression and Propensity Score Matching to Estimate the Effect of Coaching on the SAT*. *35*.

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition* (pp. ix, 334). Sage Publications, Inc.

Education in China is becoming increasingly unfair to the poor. (2021). Education in China Is Becoming Increasingly Unfair to the Poor | the Economist. https://www.economist.com/china/2021/05/27/education-in-china-is-becoming-increasingly-unfair-to-the-poor

Elder, C., & O'Loughlin, K. (2003). Investigating the relationship between intensive EAP training and band score gains on IELTS. *IELTS Research Reports*, *4*, 207–254.

Ercanbrack, J. (1999). A Study of the Effect of Direct Test Preparation on the TOEIC Scores of Japanese University Students. *Tesl-Ej*, *3*(4).

Farnsworth, T. (2013). Effects of Targeted Test Preparation on Scores of Two Tests of Oral English as a Second Language. *TESOL Quarterly*, *47*(1), 148–155. https://doi.org/10.1002/tesq.75

Filizola, E. (2008). The effect of a test preparation course on the SAT scores of students at Saint Joseph academy. *Dissertation Abstracts International Section A: Humanities and Social Sciences*, (May), 1264.

Firestone, W. A., Monfils, L., & Schorr, R. Y. (2004). Test preparation in New Jersey: Inquiry‐oriented and didactic responses. *Assessment in Education: Principles, Policy & Practice*, *11*(1), 67–88. https://doi.org/10.1080/0969594042000209001

Freda, N. M., & Lipp, M. J. (2016). Test-Enhanced Learning in Competence-Based Predoctoral Orthodontics: A Four-Year Study. *Journal of Dental Education*, *80*(3), 348–354. https://doi.org/10.1002/j.0022-0337.2016.80.3.tb06091.x

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology*, *60*(7), 991–1004. https://doi.org/10.1080/17470210600823595

Gan, Z. (2009). IELTS Preparation Course and Student IELTS Performance: A Case Study in Hong Kong. *RELC Journal: A Journal of Language Teaching and Research*, *40*(1), 23–41. https://doi.org/10.1177/0033688208101449

Gaye, Z. Z. (2001). The ACT Prep and its effects on students who participated in the Multicultural Educational Enrichment (MEE) program at Tennessee State University. Tennessee State University.

Gebril, A., & Eid, M. (2017). Test Preparation Beliefs and Practices in a High-Stakes Context: A Teacher's Perspective. *Language Assessment Quarterly*, *14*(4), 360–379. https://doi.org/10.1080/15434303.2017.1353607

Geller, J., Toftness, A., Carpenter, P., Manz, C., Coffman, C., & Lamm, M. (2017). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory*, *26*, 1–9. https://doi.org/10.1080/09658211.2017.1397175

Goossens, N., Camp, G., Verkoeijen, P. P. J. L., Tabbers, H., & Zwaan, R. (n.d.). *The benefit of retrieval practice over elaborative restudy in primary school vocabulary learning*. https://doi.org/10.1016/j.jarmac.2014.05.003

Green, A. (2007). Washback to learning outcomes: A comparative study of IELTS preparation and university pre-sessional language courses. *Assessment in Education: Principles, Policy and Practice*, *14*(1), 75–97. https://doi.org/10.1080/09695940701272880

Greving, S., & Richter, T. (2018). Examining the Testing Effect in University Teaching: Retrievability and Question Format Matter. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02412

Gribbons, B., & Herman, J. (1997). True and Quasi-Experimental Designs. *Practical Assessment, Research & Evaluation*. https://doi.org/10.7275/FS4Z-NB61

Griffin, B., Carless, S., & Wilson, I. (2013). The effect of commercial coaching on selection test performance. *Medical Teacher*, *35*(4), 295–300. https://doi.org/10.3109/0142159X.2012.746451

Gurung, R. A. R., & Burns, K. (2019). Putting evidence-based claims to the test: A multi-site classroom study of retrieval practice and spaced practice. *Applied Cognitive Psychology*, *33*(5), 732–743. https://doi.org/10.1002/acp.3507

Händel, M., Harder, B., & Dresel, M. (2020). Enhanced monitoring accuracy and test performance: Incremental effects of judgment training over and above repeated testing. *Learning and Instruction*, *65*, 101245. https://doi.org/10.1016/j.learninstruc.2019.101245

Hardison, C. M., & Sackett, P. R. (2008). Use of Writing Samples on Standardized Tests: Susceptibility to Rule-Based Coaching and the Resulting Effects on Score Improvement. *Applied Measurement in Education*, *21*(3), 227–252. https://doi.org/10.1080/08957340802161782

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review*, *19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge. https://doi.org/10.4324/9780203887332

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hausknecht, J., Halpert, J., Paolo, N., & Gerrard, M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology*, *92*, 373–385. https://doi.org/10.1037/0021-9010.92.2.373

Heitmann, S., Grund, A., Fries, S., Berthold, K., & Roelle, J. (2022). The quizzing effect depends on hope of success and can be optimized by cognitive load-based adaptation. *Learning and Instruction*, *77*, 101526. https://doi.org/10.1016/j.learninstruc.2021.101526

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta‐analysis. *Statistics in medicine*, 21(11), 1539-1558.

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (Sometimes) Enhances Learning: Performance Pressure Reduces the Benefits of Retrieval Practice. *Applied Cognitive Psychology*, *28*(4), 597–606. https://doi.org/10.1002/acp.3032

Holmes, C. T., & Keffer, R. L. (1995). A Computerized Method to Teach Latin and Greek Root Words: Effect on Verbal SAT Scores. *Journal of Educational Research*, *89*(1), 47–50. https://doi.org/10.1080/00220671.1995.9941192

Hong, E., & Peng, Y. (2008). Do Chinese students' perceptions of test value affect test performance? Mediating role of motivational and metacognitive regulation in test preparation. *Learning and Instruction*, *18*(6), 499–512. https://doi.org/10.1016/j.learninstruc.2007.10.002

Hora, M. T., & Oleson, A. K. (2017). Examining study habits in undergraduate STEM courses from a situative perspective. *International Journal of STEM Education*, *4*(1), 1. https://doi.org/10.1186/s40594-017-0055-6

Holthaus, R. A. (2008). *The effects of a high school coaching program on SAT-mathematics test scores.* Walden University.

Ireson, J. (2004). Private tutoring: How prevalent and effective is it? *Http://Lst-Iiep.Iiep-Unesco.Org/Cgi-Bin/Wwwi32.Exe/[In=epidoc1.in]/?T2000=023041/(100)*, *2*. https://doi.org/10.1080/1474846042000229458

Jäger, D. J., Maag Merki, K., Oerke, B., & Holmeier, M. (2012). Statewide low-stakes tests and a teaching to the test effect? An analysis of teacher survey data from two German states. *Assessment in Education: Principles, Policy & Practice*, *19*(4), 451–467. https://doi.org/10.1080/0969594X.2012.677803

Janes, J. L., Dunlosky, J., & Rawson, K. A. (2018). How Do Students Use Self-Testing Across Multiple Study Sessions When Preparing for a High-Stakes Exam? *Journal of Applied Research in Memory and Cognition*, *7*(2), 230–240. https://doi.org/10.1016/j.jarmac.2017.11.003

Jeynes, W. (2008). What We Should and Should Not Learn From the Japanese and Other East Asian Education Systems. *Educational Policy*, 22(6), 900–927. https://doi.org/10.1177/0895904807310042

Justus, L. C. (2010). *Impact of a school-based test preparation course on ACT scores with consideration of cultural associations.* Union University.

Kalyuga, S. (2009). *The Expertise Reversal Effect* [Chapter]. Managing Cognitive Load in Adaptive Multimedia Learning; IGI Global. https://doi.org/10.4018/978-1-60566-048-6.ch003

Kane, M. (2012). Validating score interpretations and uses. *Language Testing*, *29*(1), 3–17. https://doi.org/10.1177/0265532211417210

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Kapler, I. V., Cepeda, N. J., & Weston, T. (2012). Spacing, Testing, and Feedback: Helping Students Overcome Forgetting. *Education Canada*, *52*(1).

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486. https://doi.org/10.1037/a0017341

Karpicke, J. D. (2017). Retrieval-Based Learning: A Decade of Progress. In *Learning and Memory: A Comprehensive Reference* (pp. 487–514). Elsevier. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*(4), 471–479. https://doi.org/10.1080/09658210802647009

Khodabakhshzadeh, H., & Zardkanloo, R. (2017). The effect of mock tests on Iranian EFL learners' test scores. *International Journal of Education and Literacy Studies*, 5(3), 47-51.

Kipnis, A. B. (2011). Governing Educational Desire: Culture, Politics, and Schooling in China. In *Governing Educational Desire*. University of Chicago Press. https://doi.org/10.7208/9780226437569

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*. https://doi.org/10.1016/j.cogpsych.2019.101237

Kitsantas, A. (2002). Test Preparation and Performance: A Self-Regulatory Analysis. *The Journal of Experimental Education*, *70*(2), 101–113. https://doi.org/10.1080/00220970209599501

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, *36*(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x

Koretz, D., Mccaffrey, D., & Hamilton, L. (2001). *Toward a Framework for Validating Gains Under High-Stakes Conditions CSE Technical Report 551*.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*(9), 1297–1317. https://doi.org/10.1002/acp.1537

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219–224. https://doi.org/10.3758/BF03194055

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493–501. https://doi.org/10.1080/09658210902832915

Kuhbandner, C., & Emmerdinger, K. J. (2019). Do students really prefer repeated rereading over testing when studying textbooks? A reexamination. *Memory*, *27*(7), 952–961. https://doi.org/10.1080/09658211.2019.1610177

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*(2), 179–188.

Kulik, J. A., Kulik, C.-L. C., & Bangert, R. L. (1984). Effects of Practice on Aptitude and Achievement Test Scores. *American Educational Research Journal*, *21*(2), 435–447. https://doi.org/10.3102/00028312021002435

Kwok, P. L. Y. (2010). Demand intensity, market parameters and policy responses towards demand and supply of private supplementary tutoring in China. *Asia Pacific Education Review*, *11*(1), 49–58. https://doi.org/10.1007/s12564-009-9060-x

Kwon, S. K., Lee, M., & Shin, D. (2017). Educational assessment in the Republic of Korea: Lights and shadows of high-stake exam-based education system. *Assessment in Education: Principles, Policy & Practice*, *24*(1), 60–77. https://doi.org/10.1080/0969594X.2015.1074540

Lane, K. L., Kalberg, J. R., Mofield, E., Wehby, J. H., & Parks, R. J. (2009). Preparing students for college entrance exams: Findings of a secondary intervention conducted within a three-tiered model of support. *Remedial and Special Education*, *30*(1), 3–18. https://doi.org/10.1177/0741932507314022

Lee, J.-Y. (2019). Pedagogical Effects of Teaching Test-Taking Strategies to EFL College

Students. *Reading in a Foreign Language*, *31*(2), 226–248.

Lee, M., & Larson, R. (2000). The Korean 'Examination Hell': Long Hours of Studying, Distress, and Depression. *Journal of Youth and Adolescence*, *29*(2), 249–271. https://doi.org/10.1023/A:1005160717081

Leung, K. S. F. (2006). Mathematics Education in East Asia and the West: Does Culture Matter? In F. K. S. Leung, K.-D. Graf, & F. J. Lopez-Real (Eds.), *Mathematics Education in Different Cultural Traditions-A Comparative Study of East Asia and the West: The 13th ICMI Study* (pp. 21–46). Springer US. https://doi.org/10.1007/0-387-29723-5_2

Li, H., & Xiong, Y. (2018). The relationship between test preparation and state test performance: Evidence from the Measure of Effective Teaching (MET) project. *Education Policy Analysis Archives*, *26*, 64–64. https://doi.org/10.14507/epaa.26.3530

Li, W., & Li, Y. (2010). An Analysis on Social and Cultural Background of the Resistance for China's Education Reform and Academic Pressure. *International Education Studies*, *3*(3), 211–215.

Lipko-Speed, A., Dunlosky, J., & Rawson, K. A. (2014). Does testing with feedback help grade-school children learn key concepts in science? *Journal of Applied Research in Memory and Cognition*, *3*(3), 171–176. https://doi.org/10.1016/j.jarmac.2014.04.002

Lipman, P. (2002). Making the Global City, Making Inequality: The Political Economy and Cultural Politics of Chicago School Policy. *American Educational Research Journal*, *39*(2), 379–419. https://doi.org/10.3102/00028312039002379

Ma, H., & Chong, S. W. (2022). Predictability of IELTS in a high-stakes context: A mixed methods study of Chinese students' perspectives on test preparation. *Language Testing in Asia*, *12*(1), 2. https://doi.org/10.1186/s40468-021-00152-3

Madaus, G. F. (1988). The distortion of teaching and testing: High‐stakes testing and instruction. *Peabody Journal of Education*, *65*(3), 29–46. https://doi.org/10.1080/01619568809538611

Marton, F., & Säaljö, R. (1976). On Qualitative Differences in Learning—Ii Outcome as a Function of the Learner's Conception of the Task. *British Journal of Educational Psychology*, *46*(2), 115–127. https://doi.org/10.1111/j.2044-8279.1976.tb02304.x

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied*, *21*(4), 370–382. https://doi.org/10.1037/xap0000063

McGaghie, W. C., Downing, S. M., & Kubilius, R. (2004). What is the Impact of Commercial Test Preparation Courses on Medical Examination Performance? *Teaching and Learning in Medicine*, *16*(2), 202–211. https://doi.org/10.1207/s15328015tlm1602_14

McKenzie, J. E., Beller, E. M., & Forbes, A. B. (2016). Introduction to systematic reviews and meta-analysis. *Respirology*, 21(4), 626-637.

McMann, P. K. (1994). The effects of teaching practice review items and test-taking strategies on the ACT mathematics scores of second-year algebra students (Doctoral dissertation, Wayne State University).

Mehrens, W. A., Popham, W. J., & Ryan, J. M. (1998). How to Prepare Students for PerEorrnance Assessments. *Educational Measurement: Issues and Practice*, *17*(1), 18–22. https://doi.org/10.1111/j.1745-3992.1998.tb00617.x

Messick, S. (1982). Issues of effectiveness and equity in the coaching controversy: Implications for educational and testing practice. *Educational Psychologist*, *17*(2), 67–91. https://doi.org/10.1080/00461528209529246

Messick, S. (1989). Validity. In *Educational measurement, 3rd ed* (pp. 13–103). American Council on Education.

Messick, S. (1996). Validity and Washback in Language Testing. *ETS Research Report Series*, *1996*(1), i–18. https://doi.org/10.1002/j.2333-8504.1996.tb01695.x

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*, 707–726. https://doi.org/10.1177/001316446502500304

Mitchell, I., Nistor, N., Baltes, B., & Brown, M. (2016). Effect of Vocabulary Test Preparation on Low-Income Black Middle School Students' Reading Scores. *Journal of Educational Research and Practice*, *6*(1), 105–118.

Miyasaka, J. R. (2000). *A Framework for Evaluating the Validity of Test Preparation Practices*. The Annual Meeting of the American Educational Research Association (NewOrleans,LA,April 24-28,2000).

MOE. (2021). *National Basics*. https://www.moe.gov.cn/jyb_sjzl/moe_560/2020/quanguo/全国基本情况 - 中华人民共和国教育部政府门户网站

Montgomery, P., & Lilly, J. (2012). Systematic reviews of the effects of preparatory courses on university entrance examinations in high school-age students. *International Journal of Social Welfare*, *21*(1), 3–12. https://doi.org/10.1111/j.1468-2397.2011.00812.x

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257–271. https://doi.org/10.1080/09658211.2014.1001992

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval Practice in Classroom Settings: A Review of Applied Research. *Frontiers in Education*, *4*. https://www.frontiersin.org/articles/10.3389/feduc.2019.00005

Moore, R., Sanchez, E., & San Pedro, M. O. (2018). Investigating Test Prep Impact on Score Gains Using Quasi-Experimental Propensity Score Matching. ACT Working Paper 2018-6. *ACT, Inc.*

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257–271. https://doi.org/10.1080/09658211.2014.1001992

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Newton, P. E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education: Principles, Policy & Practice*, *14*(2), 149–170. https://doi.org/10.1080/09695940701478321

Nishitani, A. (2006). Teaching grammar for the TOEIC test: Is test preparation instruction effective?

Niu, W. (2007). Western influences on Chinese educational testing. *Comparative Education*, 43(1), 71–91. https://doi.org/10.1080/03050060601162412

Nguyen, T. N. H. (2007). Effects of Test Preparation on Test Performance －. *Melbourne Papers in Language Testing*.

Nichols, S. L., & Berliner, D. C. (2007). *Collateral Damage: How High-Stakes Testing Corrupts America's Schools*. Harvard Education Press.

O'Sullivan, B., Dunn, K., & Berry, V. (2021). Test preparation: An international comparison of test takers' preferences. *Assessment in Education: Principles, Policy & Practice*, *28*(1), 13–36. https://doi.org/10.1080/0969594X.2019.1637820

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, *38*(1), 1–4. https://doi.org/10.1207/S15326985EP3801_1

Paas, F., Renkl, A., & Sweller, J. (2004). Cognitive Load Theory: Instructional Implications of the Interaction between Information Structures and Cognitive Architecture. *Instructional Science*, *32*(1/2), 1–8.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*, 710–756. https://doi.org/10.1037/bul0000151

Pan, Y. C. (2016). Traditional and non-traditional test preparation practices: Learner performance and perspectives. *Electronic Journal of Foreign Language Teaching*, *13*(2), 170–183.

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*. https://doi.org/10.1007/s11092-016-9248-7

Peterson, D. J., & Wissman, K. T. (2018). The testing effect and analogical problem-solving. *Memory*, *26*(10), 1460–1466. https://doi.org/10.1080/09658211.2018.1491603

Phakiti, A. (2003). A Closer Look at Gender and Strategy Use in L2 Reading. *Language Learning*, *53*(4), 649–702. https://doi.org/10.1046/j.1467-9922.2003.00239.x

Phelps, R. P. (2012). The Effect of Testing on Student Achievement, 1910–2010. *International Journal of Testing*, *12*(1), 21–43. https://doi.org/10.1080/15305058.2011.602920

Pintrich, P. R. (2002). The Role of Metacognitive Knowledge in Learning, Teaching, and Assessing. *Theory Into Practice*, *41*(4), 219–225. https://doi.org/10.1207/s15430421tip4104_3

Popham, W. J. (1991). Appropriateness of Teachers' Test-Preparation Practices. *Educational Measurement: Issues and Practice*, *10*(4), 12–15. https://doi.org/10.1111/j.1745-3992.1991.tb00211.x

Popham, W. J. (2001). Teaching to the Test? *Educational Leadership*, *58*(6), 16–20.

Powers, D. E., & Rock, D. A. (1999). Effects of Coaching on SAT I: Reasoning Test Scores. *Journal of Educational Measurement*, *36*(2), 93–118. https://doi.org/10.1111/j.1745-3984.1999.tb00549.x

Rao, C., McPherson, K., Chand, R., & Khan, V. (2003). Assessing the impact of IELTS preparation programs on candidates' performance on the GT reading and writing test modules (IELTS research reports Vol. 5, pp. 237-262). Canberra: IELTS Australia.

Rawson, K. A., & Dunlosky, J. (2012). When Is Practice Testing Most Effective for Improving the Durability and Efficiency of Student Learning? *Educational Psychology Review*, *24*(3), 419–435. https://doi.org/10.1007/s10648-012-9203-1

Rawson, K. A., Dunlosky, J., & Sciartelli, S. M. (2013). The Power of Successive Relearning: Improving Performance on Course Exams and Long-Term Retention. *Educational Psychology Review*, *25*(4), 523–548. https://doi.org/10.1007/s10648-013-9240-4

Raymond, D. J. W., & A. (2021). Comparison of Traditional and Game Enhanced Test Preparation in Technical Education. Grand Canyon University.

Read, J., & Hayes, B. (2003). The impact of IELTS on preparation for academic study in New Zealand. In *International English Language Testing System (IELTS) Research Reports 2003: Volume 4* (world; pp. 153–191). IDP: IELTS Australia. https://search.informit.org/doi/abs/10.3316/informit.909013632781357

Renkl, A. (2005). The worked-out-example principle in multimedia learning. *The Cambridge Handbook of Multimedia Learning*, 229–245.

Rivers, M. L. (2021). Metacognition About Practice Testing: A Review of Learners' Beliefs, Monitoring, and Control of Test-Enhanced Learning. *Educational Psychology Review*, *33*(3), 823–862. https://doi.org/10.1007/s10648-020-09578-2

Roberts, S. O., & Oppenheim, D. B. (1966). The effect of Special Instruction upon Test Performance of High School Students in Tennessee. *ETS Research Bulletin Series*, 1966(2), i-26.

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., & Karpicke, J. D. (2006). Test-Enhanced Learning: Taking Memory Tests Improves Long-Term Retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Sadeghi, S., & Ketabi, S. (2014). High-stake test preparation courses: Washback in accountability contexts. *Journal of Education & Human Development*, *3*(1), Article 1.

Saif, S., Ma, J., May, L., & Cheng, L. (2021). Complexity of test preparation across three contexts: Case studies from Australia, Iran and China. *Assessment in Education: Principles, Policy & Practice*, *28*(1), 37–57. https://doi.org/10.1080/0969594X.2019.1700211

Samson, G. E. (1985). Effects of Training in Test-Taking Skills on Achievement Test Performance: A Quantitative Synthesis. *The Journal of Educational Research*, *78*(5), 261–266. https://doi.org/10.1080/00220671.1985.10885613

Scharfen, J., Jansen, K., & Holling, H. (2018). Retest effects in working memory capacity tests: A meta-analysis. *Psychonomic Bulletin & Review*, *25*(6), 2175–2199. https://doi.org/10.3758/s13423-018-1461-6

Scholes, R. J., & Lain, M. M. (1997). *The Effects of Test Preparation Activities on ACT Assessment Scores*. https://eric.ed.gov/?id=ED409341

Schommer-Aikins, M., & Easter, M. (2008). Epistemological beliefs' contributions to study strategies of Asian Americans and European Americans. *Journal of Educational Psychology*, *100*, 920–929. https://doi.org/10.1037/0022-0663.100.4.920

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). Meta-analysis with R: Springer.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing Effect in the Psychology Classroom: A Meta-Analytic Perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453–472. https://doi.org/10.1023/A:1003196224280

Sesnowitz, M., Bernhardt, K. L., & Knain, D. M. (1982). An analysis of the impact of commercial test preparation courses on SAT scores. *American Educational Research Journal*, *19*, 429–441. https://doi.org/10.2307/1162723

Shepard, L. A. (1990). Inflated Test Score Gains: Is the Problem Old Norms or Teaching the Test? *Educational Measurement: Issues and Practice*, *9*(3), 15–22. https://doi.org/10.1111/j.1745-3992.1990.tb00374.x

Smith, M. L. (1991). Meanings of Test Preparation. *American Educational Research Journal*, *28*(3), 521–542. https://doi.org/10.3102/00028312028003521

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784–802. https://doi.org/10.1080/09658211.2013.831454

Smolen, P., Zhang, Y., & Byrne, J. H. (2016). The right time to learn: Mechanisms and optimization of spaced learning. *Nature Reviews Neuroscience*, *17*(2), Article 2. https://doi.org/10.1038/nrn.2015.18

Stevenson, D. L., & Baker, D. P. (1992). Shadow Education and Allocation in Formal Schooling: Transition to University in Japan. *American Journal of Sociology*, *97*(6), 1639–1657.

Stobart, G., & Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1–6. https://doi.org/10.1080/0969594X.2012.639191

Stoet, G. (2010). PsyToolkit: A software package for programming psychological experiments using Linux. *Behavior Research Methods*, *42*(4), 1096–1104. https://doi.org/10.3758/BRM.42.4.1096

Stoet, G. (2017). PsyToolkit: A Novel Web-Based Method for Running Online Questionnaires and Reaction-Time Experiments. *Teaching of Psychology*, *44*(1), 24–31. https://doi.org/10.1177/0098628316677643

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction*, *4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Sweller, J. (2010). Cognitive load theory: Recent theoretical advances. In *Cognitive load theory* (pp. 29–47). Cambridge University Press. https://doi.org/10.1017/CBO9780511844744.004

Swinton, S. S., & Powers, D. E. (1983). A study of the effects of special preparation on GRE analytical scores and item types. *Journal of Educational Psychology*, *75*(1), 104–115.

https://doi.org/10.1037/0022-0663.75.1.104

Takallou, F., Vahdany, F., Araghi, S. M., & Tabrizi, A. R. N. (2015). The effect of test taking strategy instruction on iranian high school students' performance on english section of the university entrance examination and their attitude towards using these strategies. *International Journal of Applied Linguistics and English Literature*, *4*(6), 119–129. https://doi.org/10.7575/aiac.ijalel.v.4n.6p.119

Tanner-Smith, E.E., Tipton, E. & Polanin, J.R. Handling Complex Meta-analytic Data Structures Using Robust Variance Estimates: a Tutorial in R. J Dev Life Course Criminology 2, 85–112 (2016). https://doi.org/10.1007/s40865-016-0026-5

The Princeton Review. (2021). *SAT 1400 | SAT Prep Classes | The Princeton Review*. https://www.princetonreview.com/college/sat-honors-course

Thomas, R. C., Weywadt, C. R., Anderson, J. L., Martinez-Papponi, B., & McDaniel, M. A. (2018). Testing Encourages Transfer Between Factual and Application Questions in an Online Learning Environment. *Journal of Applied Research in Memory and Cognition*, *7*(2), 252–260. https://doi.org/10.1016/j.jarmac.2018.03.007

Toppino, T. C., LaVan, M. H., & Iaconelli, R. T. (2018). Metacognitive control in self-regulated learning: Conditions affecting the choice of restudying versus retrieval practice. *Memory & Cognition*, *46*(7), 1164–1177. https://doi.org/10.3758/s13421-018-0828-2

Trenkic, D., & Hu, R. (2021). Teaching to the test: The effects of coaching on English-proficiency scores for university entry. *Journal of the European Second Language Association*, *5*(1), Article 1. https://doi.org/10.22599/jesla.74

Tullis, J. G., Fiechter, J. L., & Benjamin, A. S. (2018). The efficacy of learners' testing choices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 540–552. https://doi.org/10.1037/xlm0000473

Turingan, J. P., & Yang, Y.-C. (2009). A Cross-Cultural Comparison of Self-Regulated Learning Skills between Korean and Filipino College Students. *Asian Social Science*, *5*(12), p3. https://doi.org/10.5539/ass.v5n12p3

van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology*, *36*(3), 212–218. https://doi.org/10.1016/j.cedpsych.2010.10.004

van Gog, T., & Kester, L. (2012). A Test of the Testing Effect: Acquiring Problem-Solving Skills From Worked Examples. *Cognitive Science*, *36*(8), 1532–1541. https://doi.org/10.1111/cogs.12002

van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. J. L. (2015). Testing After Worked Example Study Does Not Enhance Delayed Problem-Solving Performance Compared to Restudy. *Educational Psychology Review*, *27*(2), 265–289. https://doi.org/10.1007/s10648-015-9297-3

van Gog, T., Kester, L., & Paas, F. (2011). Effects of worked examples, example-problem, and problem-example pairs on novices' learning. *Contemporary Educational Psychology*, *36*(3), 212–218. https://doi.org/10.1016/j.cedpsych.2010.10.004

Veltre, M. T., Cho, K. W., & Neely, J. H. (2015). Transfer-appropriate processing in the testing effect. *Memory*, *23*, 1229–1237. https://doi.org/10.1080/09658211.2014.970196

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*(8), 1183–1195. https://doi.org/10.1002/acp.1630

Wang, S. (2022). *Studying China's "Super Secondary School" Phenomenon through the "Hengzhong Model"* (pp. 117–133). Brill. https://doi.org/10.1163/9789004511774_008

Wheeler, M. A., & Roediger, H. L. (1992). Disparate Effects of Repeated Testing: Reconciling Ballard's (1913) and Bartlett's (1932) Results. *Psychological Science*, *3*(4), 240–246. https://doi.org/10.1111/j.1467-9280.1992.tb00036.x

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571–580. https://doi.org/10.1080/09658210244000414

Winke, P., & Lim, H. (2017). The Effects of Test Preparation on Second-Language Listening Test Performance. Language Assessment Quarterly, 14(4), 380–397. https://doi.org/10.1080/15434303.2017.1399396

Winne, P. H., & Hadwin, A. F. (1998). Studying as Self-Regulated Learning. In *Metacognition in Educational Theory and Practice*. Routledge.

Wyn, J., Turnbull, M., & Grimshaw, L. (2014). *The impacts of high stakes testing on school students and their families*. Whitlam Institute, Sydney, Australia.

Xie, Q. (2013). Does Test Preparation Work? Implications for Score Validity. *Language Assessment Quarterly*, *10*, 196–218. https://doi.org/10.1080/15434303.2012.721423

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with Structural Equation Modeling. *Language Testing*, *30*(1), 49–70. https://doi.org/10.1177/0265532212442634

Yan, V. X., Thai, K.-P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition*, *3*(3), 140–152. https://doi.org/10.1016/j.jarmac.2014.04.003

Yang, C., Luo, L., Vadillo, M., Yu, R., & Shanks, D. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*. https://doi.org/10.1037/bul0000309

Ye, H. (2015). Key-Point Schools and Entry into Tertiary Education in China. *Chinese Sociological Review*, *47*(2), 128–153. https://doi.org/10.1080/21620555.2014.990321

Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology*, *111*, 73–90. https://doi.org/10.1037/edu0000268

Yu, G., & Green, A. (2021). Preparing for admissions tests in English. *Assessment in Education: Principles, Policy & Practice*, *28*(1), 1–12. https://doi.org/10.1080/0969594X.2021.1880120

Yu, G., He, L., Rea-Dickins, P., Kiely, R., Lu, Y., Zhang, J., Zhang, Y., Xu, S., & Fang, L. (2017). Preparing for the Speaking Tasks of the TOEFL iBT® Test: An Investigation of the Journeys of Chinese Test Takers. *ETS Research Report Series*, *2017*(1), 1–59. https://doi.org/10.1002/ets2.12145

Zhang, Y. (2011). *The Determinants of National College Entrance Exam Performance in China—With an Analysis of Private Tutoring*. https://doi.org/10.7916/D8MP5978

Zhang, Y. (2013). Does private tutoring improve students' National College Entrance Exam performance?—A case study from Jinan, China. *Economics of Education Review*, *32*, 1–28. https://doi.org/10.1016/j.econedurev.2012.09.008

Zhang, Y., Cui, C., He, Y., & Wang, L. (2022). Does private supplementary tutoring matter in Chinese students' learning of mathematics: A longitudinal study. *ZDM: The International Journal on Mathematics Education*, *54*, 3. https://doi.org/10.1007/s11858-022-01346-6

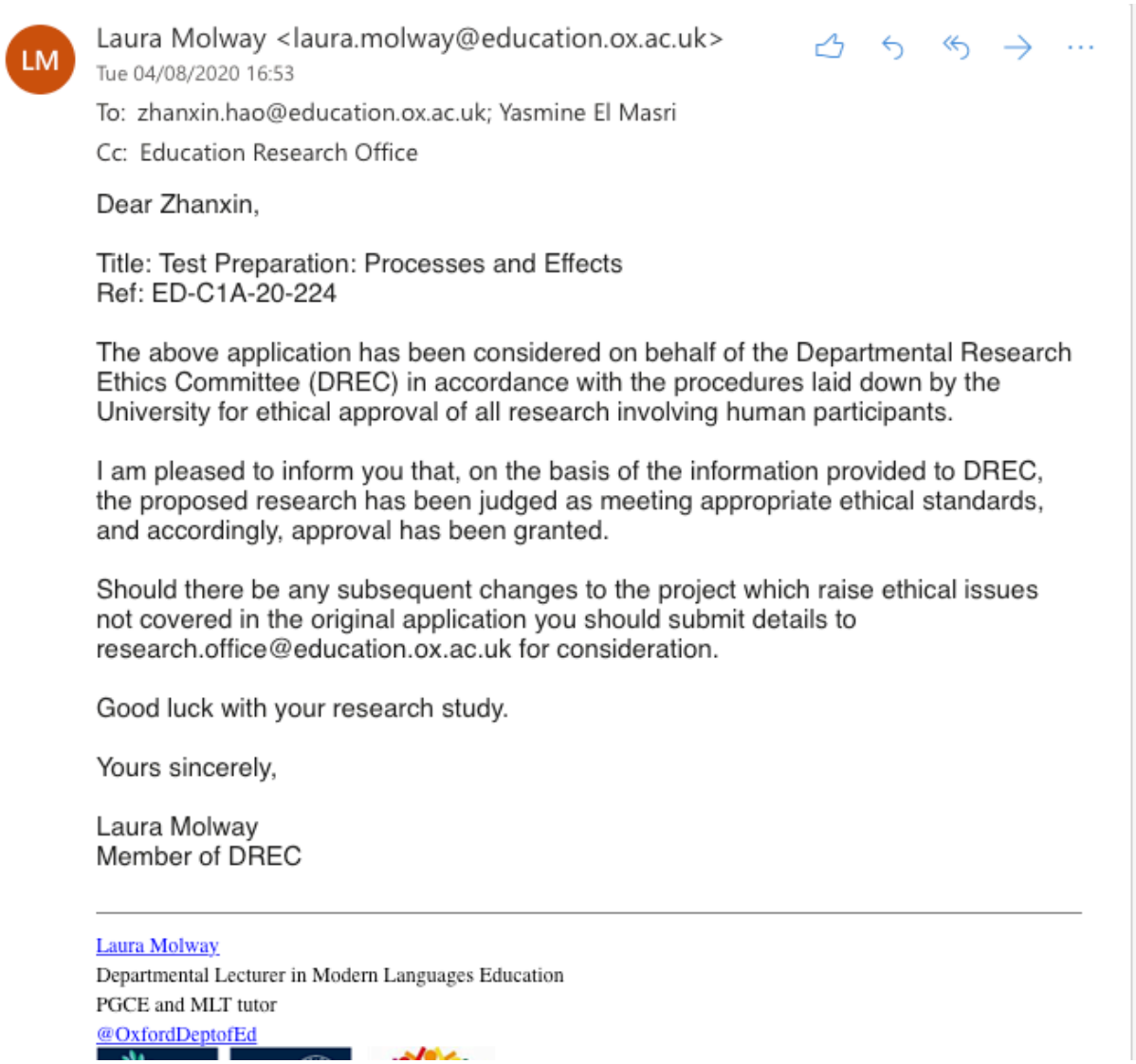
Zhou, J., & Wang, X. (2022). The Influence of Chinese College Students' Meta-Cognitive Strategy and Motivational Beliefs on their Study Habits. *The Asia-Pacific Education Researcher*, *31*(2), 93–103. https://doi.org/10.1007/s40299-020-00541-x

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7

Zuman, J. P. (1988). The Effectiveness of Special Preparation for the SAT: An Evaluation of a Commercial Coaching School.

Appendices

Appendix 1: Ethics Approval



LM

Laura Molway <laura.molway@education.ox.ac.uk>
Tue 04/08/2020 16:53
To: zhanxin.hao@education.ox.ac.uk; Yasmine El Masri
Cc: Education Research Office

Dear Zhanxin,

Title: Test Preparation: Processes and Effects
Ref: ED-C1A-20-224

The above application has been considered on behalf of the Departmental Research Ethics Committee (DREC) in accordance with the procedures laid down by the University for ethical approval of all research involving human participants.

I am pleased to inform you that, on the basis of the information provided to DREC, the proposed research has been judged as meeting appropriate ethical standards, and accordingly, approval has been granted.

Should there be any subsequent changes to the project which raise ethical issues not covered in the original application you should submit details to research.office@education.ox.ac.uk for consideration.

Good luck with your research study.

Yours sincerely,

Laura Molway
Member of DREC

Laura Molway
Departmental Lecturer in Modern Languages Education
PGCE and MLT tutor
@OxfordDeptofEd

Appendix 2: Information Sheet

DEPARTMENT OF EDUCATION

15 Norham Gardens, Oxford, OX2 6PY United Kingdom



Supervisors: Prof. Jo-Anne Baird; Dr. Yasmine El Masri

jo-anne.baird@education.ox.ac.uk; yasmine.elmasri@education.ox.ac.uk

Primary researcher: Zhanxin Hao (DPhil in Education)

+44 (0)7474139596 zhanxin.hao@education.ox.ac.uk

Participant Information Sheet

1. Invitation

You are being invited to participate in a research project. This information sheet aims to provide some detailed information about this research. Before you decide whether or not to take part, it may be important for you to understand why the research is being done and what it will involve. Please take some time to read the following information and discuss it with others freely if you wish. If you have any questions or would like more information, please do not hesitate to ask me. Thank you for reading this.

2. What is the purpose of the research?

Rooted in Chinese society's long history of using tests as the social selection mechanism, test preparation is regarded as a way to achieve success in tests. However, the effect of test preparation practices is uncertain. It has been criticized that the fever of test preparation may result in the striving for high test scores and neglecting the inherent value of what is to be learned. This research aims to investigate the effectiveness of different test preparation approaches that are commonly used by students, enlightening effective teaching and learning under the high-stakes testing agenda. Besides, students' perceptions on the test and test preparation will also be addressed in order to understand individualised test preparation pathways.

3. Why have I been selected to participate in this research?

You have been invited because you are enrolled in the specific course and preparing for the high-stakes final examination soon. Your rich experience in taking and preparing for tests throughout your academic life will be invaluable for this research.

4. Do I have to take part?

It is up to you to decide whether or not to participate. If you do decide to take part, you will be able to keep a copy of this information sheet and you will be asked to sign the consent

form to indicate your agreement. You will have the right to withdraw from the study and have your data excluded at any time without giving a reason.

5. What will happen if I take part in the research?

You will create a unique nickname or choose a given pseudonymous number before proceeding with any study. You will be asked to use this nickname or number to complete an online questionnaire first, which will take you 10-15 minutes. The questionnaire asks about your ways of learning the course materials and your methods for preparing for the final examination preparation. The questionnaire will be followed by a 30-minute pre-test. This test assesses your course-related knowledge. All test items will be in a similar format to your final course examination.

After the pre-test, you will be randomly assigned to one of the experimental groups to complete a 40-60 minutes test preparation training.

- If you are in group 1, you will do some retrieval-based learning activities, such as flashcards or quizzes, to help you learn test-related content.
- If you are in group 2, you will do same activities with group 1, but you have to answer every question correctly until you move to next activity.
- If you are in group 3, you can review the knowledge that your teacher taught in class.
- If you are in group 4, you can have some cognitive training activities such as solving puzzles and digital painting.

When you complete the training, you will need to do a post-test to see if your performance has been improved. You may also be invited to attend a follow-up interview to share more information about your learning and test preparation experience. The interview will be conducted during the course break and will be about 20 minutes long. The interview will be audio-recorded to make data collection more efficient, but you can stop the audio recording at any time.

6. What are the possible benefits of taking part?

The immediate benefit for taking part in this research is that you can receive partial course credit. No matter which group you are assigned in, as long as you complete the whole experiment (exclude the interview), you can get the credits. Moreover, you may reflect your own learning and even improve your test performance on the final course examination.

7. What are the possible disadvantages and risks of taking part?

This is a low risk research which is not anticipated to cause any disadvantage or discomfort to participants. The potential physical and/or psychological harm or distress will be the same as any experienced in everyday life. But if you feel anxious or uncomfortable during the research, you can withdraw at any time.

8. Will my data be kept confidential?

All personal data from which you might be identified (e.g. your name, your class, your college, contact details, etc.) will be kept strictly confidential. The raw information you provided will only be accessible to me and my supervisors. The data will be transferred to an anonymised form through a coding process before proceeding to analysis. You will be kept

anonymous during the data analysis and any subsequent reports or publications. All recordings will be deleted immediately after being transcribed to electronic documents. The research data will be stored confidentially using password-protected files on my encrypted personal laptop.

9. What will happen to the results of the research?

The results will inform my doctoral research. Upon successful submission, the thesis will be archived both in hard copies and an electronic version with open access. However, you and your school will not be identified in the publication. If you wish to be given an electronic copy of the thesis, please let me know.

10. Who is organising the research?

The research project is organised by Zhanxin Hao, who is a doctoral student at the University of Oxford. This project is supervised by Professor Jo-Anne Baird and Dr Yasmine El Masri from the Department of Education, University of Oxford.

11. Who has reviewed the research?

The research has been reviewed by the University of Oxford Central University Research Ethics Committee.

12. What if I have concerns about the project or I wish to complain?

If you have any concern about the project, please contact the researcher directly (please see my contact details below). I will give you response within 7 working days to indicate my intention of dealing with it. If you still have further complaints, you can contact my supervisors Jo-Anne Baird (jo- anne.baird@education.ox.ac.uk) and Yasmine El Masri (yasmine.elmasri@education.ox.ac.uk). If you feel your complaint has not been handled well, you can contact the research ethics committee (ethics@socsci.ox.ac.uk) to take your complaint further.

13. Contacts for further information

If you have any other questions or if you would like more information, please do not hesitate to let me know.

Zhanxin Hao

Doctoral Researcher
Department of Education University of Oxford
15 Norham Gardens Oxford, OX2 6PY, UK
+44 (0)7474139596
Zhanxin.hao@education.ox.ac.uk

http://www.education.ox.ac.uk/people/zhanxin-hao/

Appendix—*Consent Form*

DEPARTMENT OF EDUCATION

15 Norham Gardens, Oxford, OX2 6PY United Kingdom

Supervisors: Prof. Jo-Anne Baird; Dr. Yasmine EI Masri

jo-anne.baird@education.ox.ac.uk; yasmine.elmasri@education.ox.ac.uk

Primary researcher: Zhanxin Hao (DPhil in Education)

+44 (0)7474139596 zhanxin.hao@education.ox.ac.uk

Research Consent Form for Participants

If you are happy to participate in this study, please complete this consent form and return to the researcher.

Declaration of Consent *Please initial each box*

1. I confirm that I have read the information sheet and have had the opportunity to discuss, ask questions and receive satisfactory answers. ☐
2. I understand that my participation is voluntary, and I may withdraw at any time, without giving any reason and getting any blame. ☐
3. I understand that this project has been reviewed by, and received ethics clearance through, the University of Oxford Central University Research Ethics Committee. ☐
4. I understand that my personal data will be kept confidentially and securely in password-protected servers, and what will happen to the data at the end of the project. ☐
5. I understand that the interview will be audio-recorded but I can stop the recording at any time. ☐
6. I agree to provide my test scores, and I understand that the results will be reported pseudonymously. ☐
7. I give permission for the information gathered to be used in reporting the research findings. ☐
8. I understand how to raise a concern and make a complaint. ☐
9. I confirm that I have read the information sheet and understand the purpose of the research, my role as a participant and how data will be used in publications. ☐
10. I agree to participate in the research. ☐

Signature and Date _____

Name of Researcher_____

Signature and Date _____

Appendix 3: Test papers (exemplar)

小学教育心理学--考前强化小测试卷A

请注意：本次小测用时20分钟，请合理规划时间。请将所有答案写在答题纸上，尽量保持试卷干净整洁。答完每道题目之后，你都需要判断自己的回答是否正确，用1-10表示。1分代表“我完全不确定我的答案是正确的”，10分代表“我完全确定我的答案是正确的”。请将答案与预测值一起誊写到答题纸上。

一、单选题（10小题，每题3分，共30分。每一小题的备选答案中，只有一个答案是正确的，请把你认为正确答案的题号填入括号内。错选、多选、不选均不给分。

1. 守恒指物体不论形态如何变化，其质量恒定不变。皮亚杰认为，儿童开始具有守恒观念的年龄阶段是？（ ）

   A、0-2岁　　B、2-7岁　　C、7-11岁　　D、11-15岁

2. 既承认遗传和环境在儿童心理发展中的作用，又进一步指出，遗传对儿童心理发展作用的大小依赖于环境的变化，而环境作用的发挥，也受到遗传限度的制约。这种关于影响儿童心理发展因素的观点属于（ ）

   A、遗传决定论　　B、环境决定论

   C、遗传-环境共同决定论　　D、交互决定论

3. 根据埃里克森的人格发展理论，发展任务是获得勤奋感而克服自卑感，体验着能力的实现的年龄阶段是在（ ）

   A、幼儿阶段（0-2岁）　　B、幼儿园阶段（3-6岁）

   C、小学阶段（6-12岁）　　D、青年阶段（12-18岁）

4. 有的学生平时易受暗示，属从于权威，按照他人的意见办事，不善于适应紧急情况，这类学生的认知风格属于（ ）

   A、场依存型占优势　　B、场独立型占优势

   C、冲动型占优势　　D、沉思型占优势

5. 广义的学习是反映人和动物在生活过程中，凭借经验而产生的行为或行为潜能的相对持久的变化。下列属于学习现象的是？（ ）

   A、记住电视里战争场景发生的年份　　B、婴儿饥饿时哭泣

   C、 青春期男生的嗓音变化　　D、膝跳反射

6. 在桑代克的饿猫逃出笼子实验中，饿猫一次成功逃出获得食物的行为会得到强化，而多次出现的其他无效动作（如上蹿下跳）最终都会消失。饿猫的这种行为符合桑代克学习律中的（ ）

   A、练习律　　B、准备律　　C、效果律　　D、选择反应定律

7. 成语“举一反三”、“触类旁通"所说的学习迁移均属于（ ）

A、水平迁移 B、垂直迁移 C、负迁移 D、逆向迁移

8. 儿童刚表现改进错误的行为，家长便取消他看动画片的禁令，孩子随后表现出更多的改进错误行为。家长的行为属于（ ）

A、正强化 B、负强化 C、正惩罚 D、负惩罚

9. 学生能够运用三角形面积公式计算三角形田地的面积，这种知识属于（ ）

A、陈述性知识 B、程序性知识 C、感性知识 D、理性知识

10. 下列关于奥苏伯尔认知同化理论的阐述中，哪一句是错误的？（ ）

A、学习变化的实质在新旧知识在学习者头脑中的相互作用

B、学习的内容需要对学习者具有潜在意义和逻辑意义

C、学习的过程是学生对意义的理解过程而不是定论的传递

D、接受学习和发现学习都可以是积极主动的过程

二、名词解释题（2小题，每题15分，共30分）

1. 最近发展区

2. 社会学习理论（班杜拉）

三、分析题（2小题，每题20分，共40分）

1. 下面是两位小学学生的一次对话：

学生A: “你知道吗? 前两天在学习英语语法时，我感觉很简单，这似乎与我小时候学过一些德语有很大的关系。”

学生B: “是吗？我倒是记得在我小的时候学习拼音的拼读时，总是会跟英语字母的读法弄混。有时候会把汉语拼音中的“f”和 “g” 读成英语字母的发音，有时候又会把英语字母 “t”读成拼音的读音。”

学生A: “(若有所思)看来不同知识间的关系还是值得进一步思考的呀。”

（1） 请结合所学知识，分析材料中两位同学的谈话涉及到的学习原理 （10 分）

（2） 有哪些因素影响了这一现象的发生？（10 分）

2. 评析布鲁纳的发现学习理论（10 分）。这一理论对即将成为教师的你有什么启发？（10 分）

Appendix 4: Answer sheets (exemplar)

学生考试准备与学习策略的有效性

——基于高风险考试情境的分析

亲爱的同学:

你好，感谢你参与该项研究!

你将开始进行小测试（一），测试时长大约为20分钟。测试内容与《小学教育心理学》的课程内容和考试内容高度相关，请认真作答。

你的测试成绩仅用于本研究的数据分析，不会影响你的期末终评成绩。只有研究者本人有权限查看你的成绩和其他数据，研究者不会将你的表现对外泄漏，更不会告知你的授课老师。

在认真参与并完成全部研究流程后，你将获得平时成绩加分。如发现抄袭、敷衍等现象，将失去加分资格。

再次感谢你的配合与支持，祝你取得理想的成绩!

小学教育心理学测试（一）

答题纸

昵称：_____ 班级：_____ 试卷：A卷____B卷_____（请打勾）

一、 单选题

请在“答案”栏填写你认为正确的选项。在“判断”栏，判断你的回答是否正确，用1到10表示。1分代表“我完全不确定我的回答是正确的”，10分代表“我完全确定我的回答是正确的”。

题号	1	2	3	4	5	6	7	8	9	10
答案										
判断										
题号	11	12	13	14	15	16	17	18	19	20
答案										
判断										

二、 名词解释题

1. 请在此处回答：

请判断：你确定你的回答正确吗？请用1-10中间的数字表示，1代表完全不确定，10代表完全确定。请填写：_____

2. 请在此处回答：

请判断：你确定你的回答正确吗？请用1-10中间的数字表示，1代表完全不确定，10代表完全确定。请填写：_____

三、 论述题

1. （1）请在此处回答：

请判断：你确定你的回答正确吗？请用1-10中间的数字表示，1代表完全不确定，10代表完全确定。请填写：_____

（2）请在此处回答：

请判断：你确定你的回答正确吗？请用1-10中间的数字表示，1代表完全不确定，10代表完全确定。请填写：_____

2. **（1）请在此处回答：**

请判断：你确定你的回答正确吗？请用1-10中间的数字表示，1代表完全不确定，10代表完全确定。请填写：_____

（2）请在此处回答：

请判断：你确定你的回答正确吗？请用1-10中间的数字表示，1代表完全不确定，10代表完全确定。请填写：_____