Author:
Zhang, Ruimin

Title:
Exploring the Feasibility of Value Added Measures as an Alternative Method to Measure Public Junior High School Performance in the Context of China

Exploring the Feasibility of Value Added Measures as an Alternative Method to Measure Public Junior High School Performance in the Context of China

Ruimin Zhang

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of Doctor of Education in the Faculty of Social Sciences and Law

School of Education, June 2023

Wordcount:44,736

Abstract

With increasing international pressure to improve school education quality, accurate measures of school performance become crucial. Although raw attainment measures are commonly employed in China, value-added measures (VAMs) are considered a more scientific approach. VAMs aim to adjust for school intake differences and separate schools' influence on student outcomes from external factors (OECD, 2008). However, there is limited research on VAMs in China. This study aims to examine school and class academic performance in four student outcomes (senior high school entrance examination scores in Total, Chinese, Mathematics, and English) in Chinese public junior high schools using VAMs. Considering the challenge of implementing VAMs locally, this study also explores stakeholder perspectives on the potential benefits, disadvantages, and implementation of VAMs in the local context.

This study employs a mixed-method research design in the W district of Southwest China, including 11 public junior high schools and 46 classes. Quantitative research uses a dataset with longitudinal data to estimate student progress over time. Primary (e.g., students' questionnaires) and secondary quantitative data were collected for modelling. The analysis employs different multilevel models (raw, value-added, and contextual value-added models), all structured with three levels (student, class, school level). Adjustments for external factors demonstrate a decline in the size of school and class effects, underscoring the applicability of VAMs. Qualitative findings indicate that school evaluation results primarily serve internal accountability and improvement purposes. Policymakers show a relatively strong motivation to implement VAMs, driven by several factors. However, challenges related to other education policies, methodological concerns, and operational difficulties may hinder VAMs implementation.

Despite its limited scale, this study contributes to the methodology and practical knowledge of school evaluation in China. Unlike previous Chinese studies relying mainly on quantitative methods (Guo & Wang, 2021), this study provides richer evidence on VAMs and their implementation. However, further research with a larger sample and more diverse contexts is necessary to validate and reinforce the findings. Additionally, the insights of more practitioners are required to enhance understanding and effective implementation of VAMs.

Acknowledgements

I would like to express my deepest gratitude to all those who have supported and guided me throughout my doctoral journey and the completion of this dissertation.

First and foremost, I am immensely grateful to my supervisors, Prof. Sally Thomas and Dr. Ioanna Bakopoulou, for their unwavering guidance, invaluable expertise, and continuous support. Their mentorship has been instrumental in shaping the direction of my research and refining my academic abilities. I would like to express my special gratitude to Prof. Sally Thomas for providing me with tremendous support during the COVID-19 pandemic. Especially when I faced unexpected pressures and felt anxious, she was always there to offer timely advice and assistance. I believe that besides guiding me in my dissertation, she also patiently inspired me with valuable insights into academic research. This kind of help, which I have never received before on my academic journey, is sincerely appreciated. I am truly fortunate to have had her guidance throughout this process.

I would like to acknowledge the support provided by the staff and administrators of the School of Education. Their administrative assistance in the distant learning, extension application, and Student Visa affairs during the circumstances of COVID-19 have been indispensable in facilitating the smooth progress of my doctoral journey.

I would also like to thank the friends and colleagues, such as Yue, Qi, Agung, Mia, and Sam I met in Bristol, whose intellectual engagement, thought-provoking discussions, and encouragement have contributed immensely to my academic growth and the development of my research.

My heartfelt appreciation also goes to my family and friends for their unwavering love, encouragement, and understanding throughout this demanding endeavour. Their unwavering support, patience, and belief in my abilities have been a constant source of motivation and strength.

Lastly, I would like to express my deep appreciation to the Local Education Authority for their support in my research. Without their data support, I would not have been able to complete my study. It is through their support that my research findings have gained greater practical significance.

Overall, this dissertation would not have been possible without the collective support and encouragement of all those mentioned above. While their names may not be exhaustively listed here, their contributions are deeply appreciated and cherished.

COVID-19 Statements

The COVID-19 pandemic has had an impact on my thesis research, specifically in terms of data collection delays and the subsequent extension of my thesis timeline.

One influence of COVID-19 is the substantial delay in data collection. The restrictive measures implemented to control the spread of the virus, including lockdowns, travel restrictions, and social distancing protocols, have resulted in the delay of collecting student examination results and limited my ability to carry out the face-to-face interview. These limitations have disrupted my research plans and significantly delayed the collection of vital information required for my thesis. As a result of the data collection delays caused by the pandemic, an extension of my thesis timeline has become unavoidable. Although my own work schedule in China would be interrupted, I applied for the extension of my program. The extension allows me to interview the policymakers and headteachers in person so that more information can be obtained.

Although in the face of challenging circumstances caused by COVID-19, I have received the support from my supervisors and sought ways to adapt the unprecedented circumstances and complete my study to a required standard.

Author’s declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: Ruimin Zhang DATE: 22/06/2023

Table of Contents

List of Tables

List of Figures

Acronyms

VAMs	Value-added Measures
SER	School Effectiveness Research
ITDEQC	Improving Teacher Development and Educational Quality in China
IEEQC	Improving Educational Evaluation and Quality in China
EER	Educational Effectiveness Research
SE	School Effectiveness
SI	School Improvement
SEE	School Effectiveness Evaluation
MLM	Multilevel Modelling
VPC	Variance Partition Coefficient
SHS	Senior High Schools
JHS	Junior High Schools
LEA	Local Education Authority
VA	Value-added
CVA	Contextual Value-added
MAR	Missing at Random

Chapter 1 Introduction

1.1 Chapter Introduction

This study aims to examine the range and extent of school and class academic performance in Chinese public junior high schools using VAMs. It also seeks to explore stakeholder perspectives regarding the potential advantages, disadvantages, and implementation of VAMs to support school evaluation and improvement when the local context is considered. This chapter begins by offering a brief outline of the research problem and the rationale for undertaking new research in this area. It then outlines the research objectives and questions. Finally, the chapter concludes with an overview of the theoretical framework, the rationale for the selected methodological approach, and the organization of the thesis.

1.2 Research Background

Measuring school performance exclusively based on students' raw attainments in high-stakes examinations has been commonly used in many education systems (Leckie & Goldstein, 2017; OECD, 2008; Tang, 2005). Given the increasing international pressure to enhance the quality of school education, there has been a growing emphasis on school performance evaluation for purposes such as school accountability, improvement, and choice. The accuracy of school performance measures is crucial in achieving these objectives (OECD, 2008). However, raw attainment measures have been criticized for not adequately accounting for school intake differences (e.g., Goldstein et al., 2020; Thomas, 1998). Consequently, schools may receive lower performance measures because of their initially low-achieving intakes. This can result in various unintended consequences, such as unjust sanctions imposed on schools, misallocation of resources, and misleading feedback for school improvement (Leckie & Goldstein, 2017; Leckie & Prior, 2022; OECD, 2008; 2013; Thomas, 1998). Therefore, both researchers and policymakers have an ongoing interest in exploring scientific methods to measure school performance because valid and reliable evaluation and assessment are essential for establishing a high-performing education system.

Against this background, VAMs of school performance derived from students' scores in high-stakes examinations are widely regarded as a fairer and more scientific way to measure school performance because VAMs attempt to adjust for school intake differences and separates the contribution of schools

to student outcomes from factors that are outside the control of classes and schools (OECD, 2008). Early empirical studies conducted in the UK have shown that VAMs provide a more accurate and informative measure of school performance compared to raw attainment measures (e.g., Fitz-Gibbon, 1991; 1995 ; McPherson, 1992; Nuttall, 1991; Saunders, 1999; Thomas & Nuttall, 1993). Consequently, VAMs studies have been conducted more extensively, and VAMs have been employed in school accountability systems in some countries, including England (DfE, 2020), certain US states (Koretz, 2017), and Australia (ACARA, 2021).

In the context of China, the evaluation of public junior high school academic performance has traditionally relied on unadjusted raw attainment measures, such as average scores attained by students in the high-stake examination at the end of junior high schooling (Grade 9, age 14/15), the percentage of student achieving a score of 80 out of 100, and the pass rate (Liu & Tian, 2020) (Chinese education system will be demonstrated in Chapter 2). With the development of education in China, enhancing the quality of education has become a major priority in China, thereby resulting in a growing emphasis on reforming of school evaluation practices. Exploring new approaches to school evaluation is considered as a key aspect of improving educational evaluation, as outlined in the Overall Plan for Deepening the Reform of Education Evaluation in the New Era (State Council, 2020). Given the limitations of raw attainment measures in current school evaluation practices, new approaches to school evaluation are needed. Firstly, the new approach may provide a fairer identification of schools' contribution to student academic outcomes, serving as a vital complement to the common practice of evaluating school performance based solely on raw examination scores (Peng et al., 2006). Additionally, rather than focusing solely on raw examination scores, the development of new evaluation methods should be driven by the goal of providing more comprehensive and informative data feedback for use in policy decisions, school self-evaluation, and school improvement (OECD, 2008). Building upon previous VAMs research findings in China (e.g., Ma, 2020; Thomas, 2020; Thomas et al., 2015; Xin, 2019) and drawing from practical experiences in other countries, VAMs would provide an important addition to current school evaluation systems in China. The need for new school evaluation approaches, as outlined earlier, is further underscored by the advantages VAMs offer. Furthermore, the exploration of VAMs aligns with current education evaluation policies in China. The State Council's Overall Plan for Deepening the Reform of Education Evaluation in the New Era (State

Council, 2020) explicitly calls for exploring VAMs within the context of China. This confluence of research evidence, international experiences, and national policy direction underscores the timeliness and significance of examining VAMs in the context of Chinese education.

Moreover, it is crucial not only to determine "what works", but also to understand "how to make things work" (Hadfield & Chapman, 2015). Therefore, exploring local practitioners' perceptions on VAMs is crucial in providing evidence on the potential of implementing VAMs when considering the local context. The rationale for conducting this study will be elaborated upon in the following sections. In summary, this study seeks to produce new research evidence to inform and feed into the development of VAMs in the local context, where relevant research on VAMs is limited. It can also provide policymakers with research evidence to make informed decisions about implementing VAMs in local school evaluation practices.

1.3 Research Rationale

1.3.1 Academic Rationale

Studies in school effectiveness research (SER) emerged from debates on whether schools have an impact on student outcomes. Early contributions by Edmonds (1979), Rutter et al. (1979), and Mortimore et al. (1988) argued that schools do make a difference, challenging the conclusions of Coleman et al. (1966) and Jencks et al. (1972), who claimed minimal school effects. While historically SER has primarily been conducted in a few Western countries such as the USA (e.g., Brookover et al., 1978; Edmonds, 1979) and the UK (Rutter et al., 1979), it has now become more internationally oriented. However, given international diversity, answers to questions associated with SER tend to vary across cultures and countries (Lindorff et al., 2020). Underrepresented regions with educational systems differing from the traditional SER countries may yield contrasting findings. For example, Bosker and Witziers (1996) found larger school effects in "third world" countries compared to developed countries. Harber and Muthukrishna (2000) discovered that certain characteristics identified in traditional SER were not applicable in the South African context. Therefore, SER must be contextualized by considering social, economic, cultural, and political factors (Fertig, 2000). Conducting SER in China, an underrepresented country in the field, is meaningful. Research findings from China can contribute new evidence to validate and assess the applicability of SER theories,

methodologies, and results across different countries worldwide.

Within the context of China, previous empirical studies conducted by Thomas and colleagues (Thomas, 2005; Thomas & Peng, 2011; Thomas et al., 2012; 2015) and other more recent studies (Dong, 2021; Fan & Gao, 2019; Gao et al., 2021; Shao et al., 2021; Thomas, 2020; Zhang, 2016) have indicated that VAMs can provide greater accuracy in measuring school performance than raw attainment measures using local educational data. However, further research is necessary to examine and address certain aspects. Firstly, many previous VAMs studies in China are not longitudinal (for example, do not use prior attainment data) (Guo & Wang, 2021) and lack sufficient information on the development of the optimal model to indicate the quality of estimations. For instance, two-level value-added models (student and school levels) are commonly employed in empirical studies, yet these studies neither demonstrate the rationality of the model used nor compare different value-added models that include explanatory variables at different levels (Yang & Zhang, 2022). However, it has been claimed that the precision of VAMs results can be affected by the number of levels included in models and the explanatory variables controlled for in the analysis (e.g., Leckie & Goldstein, 2019; Mokonzi et al., 2020; Munoz-Chereau & Thomas, 2016; Thomas, 1998; Thomas & Mortimore, 1996). Therefore, it is meaningful to conduct a study that involves prior attainment data and compares different value-added models to indicate the quality of the measurement results. Furthermore, by examining the variation of student outcomes attributable to the class level, this study can enhance the comprehensiveness and accuracy of the analysis, as well as provide valuable insights for educational policy and practice.

Secondly, most quantitative research in Chinese literature has focused on outcome measures for only one or two main subjects in high-stakes examinations (Guo & Wang, 2021). However, previous studies strongly suggest the need to examine school performance across different subjects (e.g., Thomas, 1998; Thomas & Mortimore, 1996; Tymms, 1997) because important subject differences may be masked by using limited subject outcomes (Thomas & Mortimore, 1996; Yang & Zhang, 2022). Hence, it is important to employ various outcome measures to present a comprehensive picture of school performance. Overall, the points highlighted above emphasize the rationale underlying longitudinal quantitative research, which involves developing three-level models (student, class, and school levels), and comparing different value-added models across various outcome measures, to test the validity and

reliability of VAMs when the local context is considered.

Furthermore, although traditionally SER researchers have shown a strong interest in exclusively quantitative methodologies (Muijs & Brookman, 2015), there is a relative scarcity of VAMs studies in China that apply qualitative methods (Guo & Wang, 2021). This highlights a lack of research exploring the potential implementation of VAMs in the local context (Feng & Zhou, 2022). Although exclusively quantitative methods can establish the validity and reliability of VAMs, it is important not to overlook the challenges of implementing VAMs in the local context. Action-oriented research that involves qualitative methods is essential for driving changes in school performance evaluation practices. Stakeholder perceptions are critical for understanding the potential of implementing VAMs in the local context as they are familiar with the practical circumstances (Hadfield & Chapman, 2015). However, recent studies exploring the implementation of VAMs in China (e.g., Ma, 2020; Wang & Pai, 2022; Xin, 2020) predominantly focus on the perspectives of academic researchers, neglecting the viewpoints of practitioners who would be directly involved in the practical implementation of VAMs. Hence, a mixed methods design that generates both quantitative and qualitative findings is necessary to contribute original evidence to the SER field in a new context.

1.3.2 Local Rationale

Within the context of China, in addition to academic research, the Chinese central government has demonstrated its concern for developing new approaches to school performance measures, as evidenced by official policy documents that emphasize the need for educational evaluation improvement. For example, the Ninth Five-Year Plan for China's Educational Development 1995-2010 (CMOE, 2005) highlighted the necessity of developing new approaches to school evaluation and student assessment to improve educational quality. Similarly, the Opinions on Promoting the Reform of Comprehensive Evaluation of Primary and Secondary Education Quality (CMOE, 2013) emphasized the importance of considering a school's effect on student progress, rather than relying solely on students' examination results. Furthermore, the Overall Plan on Educational Evaluation Reform issued in October 2020, proposed a four-pronged evaluation approach (improving outcome evaluation, strengthening process evaluation, exploring value-added evaluation, and enhancing comprehensive evaluation) to be implemented in a coordinated manner (State Council, 2020). This

signals further government support for exploring VAMs in educational evaluation practices (Xin, 2020). Therefore, investigating VAMs aligns with the objectives of China's educational evaluation reform and has significant implications for policy development.

China's vast land area and population of over 1.4 billion results in significant diversity in economics and local cultures among different regions (National Bureau of Statistics, 2021), potentially leading to different research findings across regions (e.g., Fan & Gao, 2019; Lv, 2015; Xin et al., 2012). Despite the ongoing interest in exploring VAMs in school performance evaluation, limited research has been conducted in the Southwest region of China (Guo & Wang, 2021). Moreover, the lack of evidence on the use of VAMs in the specific local context poses a significant issue when local education authorities make policy decisions regarding school evaluation reform. Therefore, conducting a value-added evaluation study in this underrepresented area can add to the literature and provide valuable evidence.

1.3.3 Personal Rationale

My professional role and work experience have motivated me to conduct research on school effectiveness evaluation (SEE) in China. As a university teacher, I have been involved in joint projects between the university and the local education authority. During these projects, I have been conscious of the issues associated with evaluating school performance at the local government and school level. During my EdD study, I learned about value-added evaluation methods. As a critical measure of school effectiveness, it has been widely employed in many Western countries. In the context of China, a pilot 'value-added' project was conducted in China early in 2006 (Peng et al., 2006; Thomas et al., 2012); however, in the following fifteen years, this method has only been applied in a few cities. This motivates my interest in conducting in-depth research in the local context in China. Drawing upon the guidance and research evidence derived from two linked joint research projects, namely ITDEQC (Improving Teacher Development and Educational Quality in China) and IEEQC (Improving Educational Evaluation and Quality in China), this study seeks to examine the validity of the VAMs in SEE and acquire views from education practitioners about the feasibility of VAMs in practice, thereby generating new evidence for local education policy decisions.

1.4 Research Aims and Objectives

This study aims to examine the range and extent of school and class academic performance in Chinese public junior high schools using VAMs and to explore stakeholder perspectives regarding the potential advantages, disadvantages, and implementation of VAMs to support school evaluation and improvement when the local context is considered. To achieve this aim, the study aims to achieve the following objectives:

1) Review literature on SER and VAMs as well as generate a theoretical framework for this study.

2) Conduct an empirical pilot quantitative study to examine the range and extent of school and class academic performance in public junior high schools.

3) Conduct an empirical pilot qualitative study to explore the potential benefits, disadvantages and implementation of VAMs in local school evaluation practices.

4) Provide discussion and draw out implications for the methodology, policy and practice relating to VAMs and its implementation.

1.5 Research Questions

This study aims to address the following research questions (RQs):

RQ1: What is the range and extent of school and class academic performance in 11 junior high schools in the W district in Southwest China based on Raw, Value-added (controlling for prior attainment only), Contextual value-added (controlling for not only prior attainment but also student background characteristics, class, and school context) measures?

RQ2: What are stakeholders' perceptions of the potential advantages, disadvantages, and implementation of the value-added approach in school evaluation practices when local contexts are considered?

1.6 Overview of the Conceptual Framework

The integrated model of school effectiveness, originally developed by Scheerens (1990; 1992), serves as a guiding framework for conceptual modelling and variable selection of this study to address

Research Question 1 (RQ1). The model considers the school as a black box that transforms inputs into outputs through various processes. It recognizes the hierarchical structure of educational data sets, with students nested within classrooms within schools, and potentially extending to regions or nations (Creemers et al., 2010; Scheerens, 1990). By assuming that higher organizational levels facilitate effectiveness at lower levels, the model is developed and integrates various research traditions, including production functions, instructional effectiveness, and school effectiveness, to provide a comprehensive understanding of school functioning and its impact on student outcomes (Scheerens, 2005; 2013). It is important to note that the field of SER has evolved into educational effectiveness research (EER), which encompasses a broader scope beyond the study of individual schools. EER now examines the questions about the effects of teachers, classrooms, schools, and systems, to gain a comprehensive understanding of educational effectiveness (Lindorff et al., 2020). This shift further reflects a recognition that multiple levels and dimensions of the education system contribute to overall educational effectiveness. Regarding the variable selection, studies (e.g., De Jong et al., 2004; Kyriakides, 2005; Reezigt et al., 1999; Reynolds et al., 2014) have shown the influence of variables on student achievement to be multilevel. Thus, key variables from previous research, proven to be statistically significant to student outcomes, are selected and placed within the appropriate level of school functioning (e.g., Chapman et al., 2015; Hall et al., 2020; Hu et al., 2022; Mokonzi et al., 2020; Thomas, 1998; 2020; Shao et al., 2021). In Chapter 2, a detailed demonstration will be provided to indicate how the integrated model of school effectiveness informs the value-added modelling and its application in addressing RQ1.

The potential of implementing VAMs is informed by holistic approaches (Hadfield & Chapman, 2015) and previous research (e.g., Bee, 1973; Feng & Zhou, 2022; Munoz-Chereau, 2020; OECD, 2009; 2013; Peng et al., 2013; Sammons et al., 1998; Saunders, 2000; Thomas, 2020) that has examined the development and implementation of VAMs. These studies provide a foundation by creating an analytical framework for the qualitative study. Hadfield and Chapman (2015) summarise three areas that can assist researchers and practitioners in generating data, insight, analysis, and learning. The first area is to disrupt the practitioners' existing view of their own practice, encouraging them to critically reflect on their current approaches. The second area is to help practitioners understand about how the current situation has arisen and consider potential areas for change. The third area is to assist

practitioners in contextualizing new knowledge and think about effective ways to bring about change while considering the link between changes and outcomes. Accordingly, the interview schedule in this study includes stakeholder perceptions of the purpose of school evaluation, strengths and weaknesses in current school evaluation practices, the concept of value-added, motivations of stakeholders to implement VAMs, and the potential factors that help or hinder the implementation of VAMs. By addressing these aspects, the study aims to provide valuable insights into stakeholder perspectives on VAMs and their practical implementation.

1.7 Overview of Methodological Approach

To achieve the goals of this study, a pragmatist philosophical standpoint was adopted, which emphasizes practical application and uses appropriate methods to address the specific RQs that align with the researcher's values (Creswell, 2003; Tashakkori & Teddlie, 1998). Instead of replacing one another, this standpoint recognizes the advantages of both quantitative and qualitative research methods and seeks to leverage them effectively (Johnson & Onwuegbuzie, 2004). Pragmatism has been identified as a suitable paradigm for mixed-method designs, allowing for flexibility in addressing multiple research questions and employing various quantitative and qualitative analysis tools (Creswell & Clark, 2011; Teddlie & Tashakkori, 2009). In accordance with the chosen mixed-methods research design, student questionnaires were used to collect data for developing value-added models, which were then used to estimate school and class value-added performance. Qualitative semi-structured interviews were conducted to gain deeper insights into the perceptions of local policymakers and headteachers regarding the potential benefits, disadvantages and application of VAMs. The data analysis encompassed multilevel modelling techniques (Goldstein, 1997; 2011) for measuring value-added in the quantitative phase, while thematic analysis was used to explore the potential of applying VAMs in local school evaluation practice in the qualitative phase (Braun & Clarke, 2006). Further details on the research design, data collection methods, and data analysis can be found in Chapter 3.

1.8 Thesis Outline

This study is structured into seven chapters. In this introductory chapter, the research problem is briefly introduced, and the research aims and objectives are presented. The chapter also provides insights into the academic, local, and personal motivations behind conducting this research. Furthermore, it offers

a concise overview of the study's conceptual framework and research design.

Chapter 2 is a literature review on developing research focuses, conceptual frameworks, methodologies, and some criticisms of SER. It also clarifies what value-added evaluation means and how to evaluate school performance. Following a review of the literature on SEE in China, the gap in the literature addressed in this study is identified. The final section of this chapter involves the discussion of the research context.

Chapter 3 presents details of the research methods used in this study. It first identifies the philosophical standpoint of pragmatism and justifies the combination of quantitative and qualitative methods used in this study. It then describes and explains the decisions made in research sampling, data collection methods and procedures, designing and piloting the survey and interview instruments, and data analysis. The final part of this chapter demonstrates how this study's ethical issues were addressed, how validity in the research design was ensured, and potential methodological limitations.

Chapter 4 presents quantitative findings to address RQ1 and sub-questions 1.1-1.6.

Chapter 5 provides the qualitative findings to address RQ2 and sub-questions 2.1-2.5.

Chapter 6 discusses the research findings that emerge from Chapters 4 and 5, particularly in light of previous international and local research findings and the research context in China.

Chapter 7 gives this study’s conclusion, including the original contributions, the implications for policy and practice, limitations, and suggestions for further research.

Chapter 2 Literature Review

2.1 Chapter Introduction

This chapter provides a critical review of the relevant literature related to SER and VAMs of school effectiveness. The first two sections present an overview of SER and a theoretical framework for this study. Subsequent sections discuss VAMs of school effectiveness, including the concept of VAMs, and how VAMs have been developed to measure school and class performance. Chinese-based research on VAMs is also reviewed and linked to the justification for conducting new research in China. To ensure a comprehensive understanding of the research context, this chapter also presents key information about China's education system. This includes an overview of the stages of schooling, high stakes examinations, and education governance. The final section builds upon the arguments presented in the previous sections, offering a justification for the study and outlining the specific research questions of this study.

2.2 Overview of EER

2.2.1 Emergence of SER

The emergence of SER was driven by debates associated with the role of schools in shaping student outcomes. Earlier studies (e.g., Coleman et al., 1966; Jencks et al., 1972; Plowden, 1967) suggested that the impact of the students' socio-economic background and ability outweighed the effects of schooling. In response, SER emerged and challenged this pessimistic view by demonstrating that schools do make a difference in student outcomes (e.g., Brookover et al., 1979; Edmonds, 1979; Rutter et al., 1979; Weber, 1971). SER not only established the argument that schools have effects on student outcomes but also identified certain characteristics associated with successful schools (e.g., Mortimore et al., 1988; Murnane, 1981; Wynne, 1981). As SER evolved over time, it encompassed three major strands of research: school effects research, effective school research, and school improvement research (Reynolds & Teddlie, 2002). These strands demonstrate an increasing interest in not only evaluating school's capacity to enhance student achievement, as evidenced by the focus on VAMs in school performance in this study, but also in understanding how schools can be made more effective (Chapman et al., 2015). Overall, the emergence of SER represents an important development in

educational research and practice, reflecting a growing recognition of the educational system's potential to improve outcomes for all students. The following sections will highlight three key aspects of this development.

2.2.2 Evolution From SER to EER

The evolution from SER to EER reflects a shift in focus from solely examining school-wide factors to considering multiple levels and interactions in the education system. Early SER locked themselves into an almost exclusive concern with the school-wide factors rather than with the class and the teacher (Chapman et al., 2015), while teacher effectiveness studies primarily concerned with the process variables with the exclusion of school-wide factors (Teddlie, 1994). These were criticized for separately studying on school and teacher effectiveness, because neither level can be adequately studied without connecting the other (Kyriakides, 2005; Opdenakker & Van Damme, 2000; Reynolds et al. 2014). Therefore, joint studies that consider both teacher and school influences (e.g., Creemers, 1992; Mortimore et al., 1988; Teddlie & Stringfield, 1993) have been identified as a significant development in EER.

Beyond the factors associated with teachers and schools, the role of educational systems is also acknowledged as an influential element in relation to student outcomes (Scheerens, 2013). EER goes beyond examining factors at teacher and school levels, and encompasses teacher effectiveness, school effectiveness, and system effectiveness, with additional levels, such as district/local authority and regions, to explain variations in student outcomes (Creemers, 1994; Munoz-Chereau & Thomas, 2016).

The evolution from SER to EER informs the design and analysis of studies in EER, ensuring that the multi-level nature and interactions between levels are appropriately accounted for. It also highlights the significance of integrated conceptual modelling of school effectiveness, which considers the various levels and their interactions. More details will be discussed in Section 2.3.

2.2.3 Links Between School Effectiveness (SE) and School Improvement (SI)

During the 1990s, there was a limited amount of collaboration between studies within the SE and SI (Mortimore, 1998; Reynolds & Stoll, 1996). This lack of collaboration could be attributed to their different orientations and methodological approaches (Reynolds et al., 1996). SE primarily focused on

quantitative methods and school-level analysis, while SI emphasized qualitative methods and practitioner-level analysis. The SE studies tended to have a static orientation, primarily describing the contributions schools make, while SI studies took on a dynamic orientation, concerned with the practical application of research findings (Reynolds, 1988; Reynolds et al., 1996; Scheerens, 2013).

Despite these differences, both areas share the common goal of improving student outcomes, leading researchers to propose their merger (e.g., Creemers & Reynolds, 1990; Mortimore 1991; Reynolds et al., 2020; Stoll, 1996; Stoll & Fink, 1992). The knowledge gained from SE can inform SI practices by identifying characteristics and processes associated with positive student achievement (Mortimore, 1998). On the other hand, SI can test SE theories by implementing practical changes and measuring their impact on student outcomes (Teddlie & Reynolds, 2002). Along with the voices calling for improving links between SE and SI, both areas have grown together over time (Lindoff et al., 2020).

Thus, the importance of improving practices has become increasingly recognized, and the knowledge provided by SER contributes relevant evidence in answering the question of how to improve practices. Therefore, the links between SE and SI highlight the importance of not only generating empirical data for academic purposes but also considering how to improve practices for the benefit of students.

2.2.4 International Dimension of SER/EER

SER initially emerged and developed primarily in a few Western countries like the USA and the UK (Reynolds, 2002). However, with the increasing internationalization of education policy and practice, SER/EER has become an increasingly global research field. This trend includes a focus on examining educational effectiveness beyond the original contexts where it was developed (Lindorff et al., 2020).

Cross-country research has demonstrated that simply transplanting knowledge from one cultural context to another may not yield effective results (Reynolds, 2002). For example, despite extensive empirical evidence supporting the importance of assertive principal instructional leadership in the US (Levine & Lezotte, 1990), this factor was found to be invalid in the Dutch educational and social climate (van de Grift, 1990). Moreover, Heyneman and Loxley (1983) found that the proportion of variance in student achievements attributed to schools was larger in developing countries than in developed ones. These diverse research findings across different countries underscore the need for

further investigation to provide robust evidence that explains how educational concepts and factors are implemented in different country contexts (Lindorff et al ,2020). Thus, the growing international dimension of SER/EER motivates this study to be conducted in the context of China, where SER/EER is not extensively developed, and to inform evidence-based and appropriately contextualized educational practices.

2.2.5 Criticisms of SER

Based on the above review of the development of SER, it is evident that SER has significantly influenced educational research. It has created a key paradigm and methodology for measuring school effects, emphasized student outcomes as the primary goal of schooling, provided evidence on schools' effects on student outcomes, and sought to identify key school factors that enable schools to add value to their students' learning outcomes (Chapman et al., 2015). Moreover, the development of the international dimension has enabled SER to expand beyond its Western origins to diverse country contexts, particularly in developing countries where the SER knowledge base is limited (Hall et al., 2020). This expansion has allowed SER to contribute to educational research on a global scale. Despite these contributions, the development of SER has not been without criticism from various perspectives.

Firstly, SER has been criticized for its lack of consideration of social class-related issues. Thrupp (1999; 2001) argued that SER oversimplifies the issue by overemphasizing the influence of schooling while neglecting the impact of social class on learning and teaching in schools. Thrupp (2007) later advocated for SER to acknowledge the diverse local and political contexts of schools, such as student intake characteristics and school and area characteristics. Thus, it is crucial to consider the socio-economic position of students when making decisions for school improvement. In response to these criticisms, SER has evolved to incorporate more advanced methods of analysis, such as multilevel modeling (MLM), which seeks to account for various contextual factors and adjusts for factors beyond the control of schools to separate school effects from student intake effects (e.g., Goldstein, 1997; Creemers et al., 2010). However, it should be noted that MLM has its limitations when measuring school effects, as the measures used can often be crude and unable to account for significant unmeasured factors.

The other limitation of SER may be also associated with its lack of a theoretical basis. This criticism

arises from the reliance on statistical rationality to justify the inclusion of variables in models, rather than grounding them in theoretical frameworks. Consequently, correlational studies examining the relationship between school characteristics and attainment have failed to provide clear explanations as to why specific factors are expected to be associated with higher school attainment (Coe & Fitzgibbon, 1998). Scheerens and Bosker (1997) argued that school effectiveness researchers tend to report only statistically significant factors, leading to a neglect of non-significant factors. In some cases, these neglected factors may be potentially critical for education (Luyten et al., 2005).

Another criticism of SER concerns its political-ideological focus. SER research has been criticized for reflecting governmental concerns and failing to maintain objectivity. Scholars (e.g., Ball, 1998; Grace, 1998; Rea & Weiner, 1998) argued against the assumption of the SER tradition that it generates objective knowledge through rigorous quantitative methodologies. In their views, research, to some extent, involves ideological and political choices that serve specific interests. Thus, in the 1990s', SER was seen as dominant within certain governments or government-related institutions, those studies are likely to reflect government concerns rather than scientific considerations. For example, Goldstein and Myers (1997) argued that SER has been viewed by many politicians as a legitimation device for new policies such as raising standards and achieving targets. In addition, Townsend et al. (2015) discussed that SER provided measurement tools for policymakers to compare schools through performance tables. However, it may narrow the vision of what schools are because it blames schools for failing on the assumption but ignoring complex educational issues. Nevertheless, it should be recognized that debates surrounding SER/EER would always exist, and a final solution may not be attainable. However, a better understanding of the viewpoints of the critics is necessary to develop a more comprehensive reframing (Townsend et al., 2015).

Overall, reviewing the development of SER/EER provides foundational knowledge for the aim and research questions of this study. To address the research questions effectively, it is important to first gain a comprehensive understanding of VAMs and their application in measuring educational effectiveness. This be discussed in the next section.

2.3 The Conceptual Model of This Study

During the early phase of SER, the primary focus was on establishing statistical relationships between

variables. However, this has been criticized for the lack of rational models that can serve as a basis for theory-building (Creemers, 2006; Kyriakides, 2005; Scheerens, 2013). To advance the development of theoretical models and explore effective methods of measuring school effects and identifying key factors that explain variation in student outcomes, researchers in SER have sought to model school effectiveness (Creemers, 1994). These models play a crucial role in guiding the design of empirical studies within the field of SER (Kyriakides et al., 2000).

For this study, the integrated model of school effectiveness originally developed by Scheerens (1990; 1992) serves as the conceptual framework to guide the research design. This section will review relevant literature to demonstrate how the integrated model has developed and why it is employed as the guiding theoretical framework for this research.

2.3.1 Theoretical Models from Disciplinary Perspectives

2.3.1.1 Educational Production Function

In the field of SER, several perspectives attempt to explain the factors that contribute to school effectiveness. One of these perspectives is the economic approach. This approach aims to establish a function that indicates the relationship between school inputs and outcomes, while controlling for the influence of background variables (Monk, 1992). This function, known as the educational production function, was utilized in early SER. It assumes that an increase in inputs leads to an increase in outcomes (Brown & Saks, 1986). The main characteristics of this approach include the selection of resource input (e.g., student/teacher ratio, teacher salary) as the primary antecedent conditions, the measurement of direct effects, and the use of aggregated data at a single level, such as the school level or student level (Creemers, 2005).

However, some researchers have highlighted weaknesses in these models. For example, these models express the value of inputs and outputs in terms of economic resources, and Creemers and Kyriakides (2016) pointed out that determining the monetary value of inputs, processes, and outputs is challenging. Moreover, defining desired outputs presents a challenge. Furthermore, the major weakness of these models is the use of aggregated data, as it hinders the understanding of procedural and organizational measures that are essential for influencing student outcomes (Scheerens, 2013).

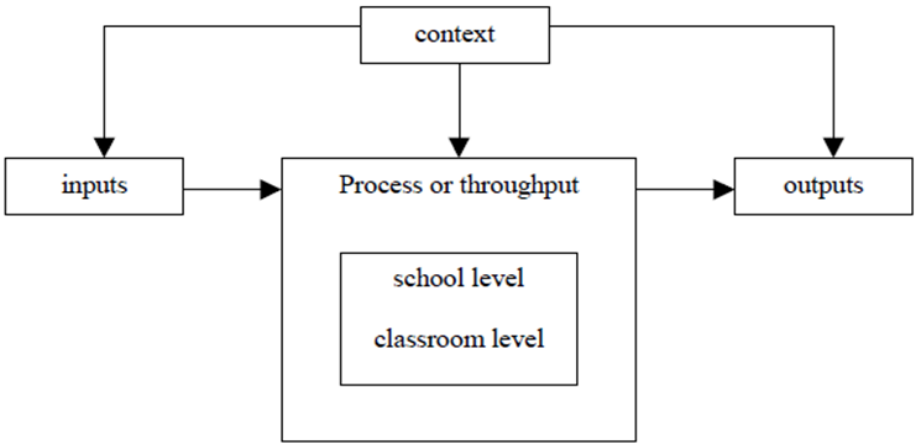
2.3.1.2 Instructional Effectiveness Model

The instructional effectiveness model is an approach to school effectiveness that focuses on context and process variables from both sociological and psychological perspectives. The sociological perspective focuses on individual background and context variables, such as student socio-economic status, gender, age, and ethnicity, while the psychological perspective links student background factors to their learning aptitude or motivation (Kelly, 2012). Variables that measure the learning processes, such as quality of instruction, opportunity to learn, and ability to understand instruction, are also concerned in these models (Scheerens, 1997). A primary characteristic of these models is their focus on measuring teacher effectiveness (Creemers & Kyriakides, 2008; Muijs et al., 2014) and guiding effectiveness research toward a more focused approach on the process occurring in the classroom (Creemers et al., 2010).

In short, various theoretical models have been developed within the field of EER based on different theoretical origins (Creemers & Kyriakides, 2016) or strands of research (Scheerens, 2005). However, all these models share a common objective of breaking open the "black box" of the school, which is the place where processes take place to transform inputs into outputs, taking into account contextual conditions (Scheerens, 2013). Therefore, enhancing school output measures, particularly student achievement, it is necessary to view schools as organizations rather than solely focusing on individual learning or teaching (Creemers & Kyriakides, 2016). A basic system model (see Figure 1) can be used to illustrate how schools function as organizations. Based on this model, integrated models have been developed to attempt to integrate the theories and findings of various approaches to educational effectiveness.

Figure 1

A basic system model of school functioning



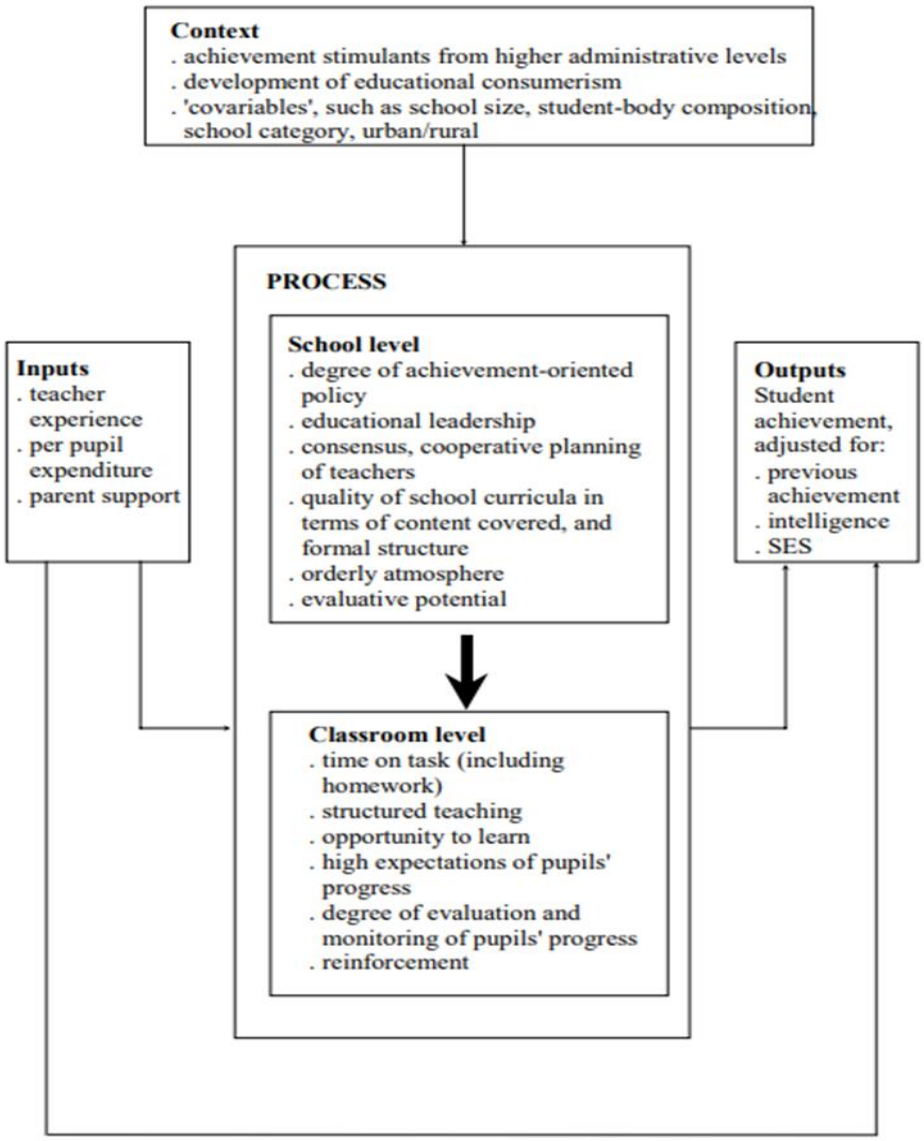
Note. The image was created by Scheerens, J. (2013). Scheerens, J. (2013). What is effective schooling? A review of current thought and practice. International Baccalaureate Organization. *Zugriff am*, 3, 2014.

2.3.2 Integrated Model of Educational Effectiveness

The integrated model of educational effectiveness is characterized by its multi-level structure, which recognizes that students are nested in classrooms, classrooms are nested in schools, and schools are nested in contexts. The model emphasizes that higher levels provide the necessary conditions for enhancing effectiveness at lower levels (Scheerens & Creemers, 1989). This allows for a synthesis of research findings from production function research, instruction effectiveness research, and school effectiveness research (see Figure 2).

Figure 2

An integrated model of school effectiveness (from Scheerens, 1990)



Note. The image was created by Scheerens, J. (1990). School effectiveness research and the development of process indicators of school functioning. *School effectiveness and school improvement*, 1(1), 61-80.

The integrated model of educational effectiveness includes a range of key variables from different traditions, with each variable located at the appropriate level of school functioning (e.g., the school, the class, and the student levels). Input variables include student background characteristics, school resources, and school context factors. Process variables are those that take place in the classroom and include teacher characteristics, teaching strategies, and student learning strategies. Output variables measure student achievement and other outcomes such as school completion and post-secondary education attainment adjusted for relevant student background factors (Scheerens, 2005).

The choice of key variables is supported by the meta-analysis and the re-analysis of selected key SER studies (Scheerens, 2005). Several exemplary studies have used this conceptual framework to measure and analyse school effectiveness, such as Hill et al. (1995), Mortimore et al. (1988), Sammons et al. (1995), and Thomas (1998). Many Chinese researchers on SER and SEE studies have also adopted this framework, including Du and Yang (2011), Fan and Gao (2019), Peng et al. (2006), Shao et al. (2021), Tang (2005), Thomas (2005; 2020), Thomas et al. (2012), and Wang et al. (2009).

In short, test-based accountability policies and school performance feedback rely on output measures, particularly student academic achievement, to evaluate school quality (Scheerens, 2005). When comparing school performance, the focus is on the instrumental value of input and process indicators to maximize output rather than organizational arrangements or teaching strategies (Scheerens, 2013). Therefore, the question of 'what works best' may refer to the extra value that school create in terms of school output. Researchers have defined this extra value as value-added (Scheerens et al., 2003). The concept of value-added and the development of value-added modelling will be discussed in detail in the following sections.

2.4 Value Added Measures of School Effectiveness

2.4.1 Defining Educational Quality and School Effectiveness

When examining the measurement of school effectiveness, it is essential to first define what are meant by educational quality and school effectiveness. Defining educational quality primarily focuses on student outcomes and two principles are commonly used to define it. The first highlights the importance of students' cognitive development as the primary objective of education systems. The second principle emphasizes the development of values, attitudes, and skills related to responsible citizenship, creativity, and emotional growth (UNESCO, 2005). Both principles acknowledge the multifaceted nature of educational quality, encompassing not only academic achievement but also broader aspects of personal, social, and emotional growth. In SER, cognitive outcomes, mostly in terms of achievement in core-subjects, have predominated (Scheerens et al., 2003).

School effectiveness generally refers to the extent to which a school achieves its educational goals and objectives (Scheerens et al., 2003). It indicates how well a school provides quality education to its

students, given the available resources and constraints. In this study, school effectiveness can be understood as an aspect of school quality, where the level of school output is at the core of quality judgement on schools (Scheerens, 2013).

There are several reasons for the measurement and evaluation of school effectiveness. These include addressing economic and social challenges, monitoring school performance, holding schools accountable, facilitating internal self-evaluation and improvement, making evidence-based decisions, and enabling school choice (OECD, 2008; Scheerens et al., 2003). However, a critical issue that has received significant attention within the field of SEE is how to accurately measure school effectiveness (Munoz-Chereau et al., 2020; OECD, 2008; 2013). Therefore, this review will focus on examining the measurement of school effectiveness, the emergence of VAMs as a valid and reliable method for evaluating school performance, and the technical aspects associated with VAMs.

2.4.2 Limitations of Raw Student Attainment Measures

In the past, school performance has commonly been evaluated based on raw unadjusted student attainment, which provides information on the level of student achievement at a specific point in time. This measure typically involves assessing the average score on standardized tests at the end of the school year, or the percentage of students progressing to higher levels of education (Leckie & Goldstein, 2017; OECD, 2008). These raw achievement scores are typically aggregated to the school level and used to rate school performance, such as through the creation of "league tables" in the UK during the 1990s and early 2000s (Bosker & Witziers, 1996).

However, the use of these raw attainment measures has faced significant criticism from both practitioners and researchers, particularly when it was used to rate school performance. Thomas (1998) argued that although raw attainment measures may indicate students' academic achievement, they do not accurately represent the effectiveness of schools for their students. Munoz-Chereau and Thomas (2016) suggested that raw measures of students' academic outcomes are insufficient for school accountability and improvement since they fail to account for contextual and external factors that influence student results. These external factors, such as students' prior attainment, student background characteristics, and contextual factors, may influence student achievement in ways that are beyond the control of schools (e.g., Crawford et al., 2007; Easen & Bolden, 2005; Haworth et al., 2011; Ray, 2006).

Therefore, solely holding schools responsible for raw results that are out of the control of schools or unevenly distributed between schools is considered inaccurate and unfair. As a result, many studies (e.g., Goldstein, 1993; McPherson, 1992; Nuttall, 1991; Raudenbush & Willms, 1996; Stoll & Mortimore, 1995) have argued for the use of more sophisticated and longitudinal data to measure school performance.

2.4.3 Definition and Advantages of VAMs

In response to the limitations of using unadjusted student raw attainments to evaluate school performance, VAMs have emerged as a more reliable and valid measure for school performance and have been recognized as the key outcome measures in SER (Scheerens et al., 2003). The key assumption of VAMs is that schools contribute value to their students' academic achievement (Goldstein, 1991; 1995; 1997; 2003). According to OECD (2008), "the term value added refers to the extra value that school contribute to students' progress towards stated education objectives (e.g., student academic achievement in this study). The contribution is net of other factors that contributed to students' educational progress" (OECD, 2008, p.17). Then, "value-added measures of school effectiveness seek to evaluate the relative progress made by students in a particular school over a specified period of time in comparison to students in other schools in the sample" (Thomas, 1998, p.94).

The literature highlights two key advantages of VAMs. First, VAMs generate more accurate school performance measurements compared to raw attainment measures. In most school settings, there are differences, such as individual student characteristics and students' prior attainment. The effects of those factors on student achievement may potentially confound the effects associated with classes or schools (Leckie & Goldstein, 2017; Raudenbush & Bryk, 2002). Thus, a straightforward comparison of school performance without considering these factors would not be like with like comparisons (Stoll & Mortimore, 1995). The accuracy argument states that schools should be held accountable only for the things they can influence (Nuttall, 1991; OECD, 2008). VAMs attempt to address the issue of possible confounding effects by adjusting for school intake differences to generate a fair and valid assessment of school performance (Sammons, 1999). VAMs are widely accepted as a more accurate tool for identifying good practice in the educational system and have played a critical role in many

school accountability systems worldwide (Leckie & Prior, 2022; OECD, 2008). However, the momentum for VAMs has not been reached in China probably associated with the relative scarcity of empirical research in the local context (see section 2.6.1). This study therefore seeks to explore school and class effects and provide new original evidence within mainland China.

Secondly, the link between SE or SEE and SI highlights that providing valid evaluation results is not the end goal; rather, valid information can be used as a basis for action that advances school improvement objectives (Thomas, 1998). For example, VAMs can identify high- or low-performing schools, inform resource allocation and teaching practices, and evaluate the impact of school improvement strategies (OECD, 2008; Sammons, 2010; Thomas, 1998). However, it should be noted that VAMs have the advantages of promoting SI relies on practitioners who can influence processes and outcomes (Hadfield & Chapman, 2015; OECD, 2008). This study, therefore, also seeks to explore the perspectives of stakeholders in Chinese junior high schools regarding the potential benefits, disadvantages, and implementation of VAMs for evaluating school performance (see section 2.6.3). The next section will review and discuss how value-added statistical models have developed to estimate the contributions of schools and classes to student progress.

2.4.4 Design of Value-Added Models

Value-added modelling refers to a set of multilevel statistical models (details will be demonstrated in Section 2.4.5) that are designed to estimate the proportion of the variance in student achievement attributed to a school over at least two points in time (Sammons et al.,2016). Raw unadjusted attainment measures fail to account for the confounding effects of school intake differences, which may generate biased school comparisons. To overcome this issue, most value-added models employ a regression adjustment to student achievement, which allows for the contribution of individual schools to their students' academic progress to be isolated (OECD, 2008). There are several types of value-added models, each grounded on the principle that student achievement is influenced by various factors operating at different levels (Keeves et al., 2005).The different models are distinguished by the various adjustments made to estimate the value-added and lead to different predicted school and class effects (Timmermans et al., 2011). In this study, four main models (raw model, value added model, contextual value-added model 1, and contextual value-added model 2) were developed to measure school and

class performance demonstrated in the following sections. It is also meaningful to compare each model to explore an optimal model in the specific context of this study.

2.4.4.1 Raw Model Makes No Statistical Adjustments

The simplest model used is the 'Raw' unadjusted model (Timmermans et al., 2011), which essentially measures the disparity between a school's performance and the average performance of students across all schools. This model provides a measure of gross school effects, representing the deviation from the overall mean (Sammons et al., 2016). However, this model does not account for any of the variation in student intake or contextual factors among schools (Leckie & Prior, 2022).

2.4.4.2 Value-Added Model (VA Model) with Adjustment for Prior Attainment

The key issue to be discussed in value-added modelling is determining which factors should be adjusted for isolate the contribution of schools to student academic performance. This is based on estimating the relationship between student academic performance and various factors (OECD, 2008). One crucial factor is student prior attainment, which is widely accepted as the most important predictor of student current achievement (Leckie & Goldstein, 2017). Therefore, the VA model employs student prior attainment as a basis for taking account of differences in enrolled students among schools. It essentially measures how much better or worse a school performs in terms of the mean increase in student learning over a period relative to the average school (Goldstein, 1997; Raudenbush & Willms, 1996). By taking the differences in student prior attainment into consideration, a considerable proportion of the variance in students' raw attainment may be explained, and controlling for student prior attainment is considered a critical factor in measuring school value added performance (Lenkeit, 2013; Teddlie & Reynolds, 2002; Thomas, 2001; Thomas et al, 1997) (see Section 2.4.4.4).

2.4.4.3 Contextualised Value-Added Models (CVA Model) with Adjustment for Context Characteristics

Despite the advantages of VA model, there are arguments against only controlling for prior attainment. Critics emphasize the importance of adjusting for relevant student background and school contextual factors when estimating school performance. Thus, the VA model is extended to the CVA model that additionally controls for intake differences in students' sociodemographic characteristics and

contextual factors (Leckie & Goldstein, 2019; Leckie & Prior, 2022).

CVA models include contextual factors because these factors can predict student achievements, vary across schools at intake, and are argued beyond the school's control (Ballou et al., 2004; Leckie & Goldstein, 2019; OECD, 2008; Raudenbush & Willms, 1996). Studies such as those by Goldstein (1997), Leckie and Goldstein (2019), Raudenbush and Willms (1996), Reynolds et al. (2014), and Teddlie and Reynolds (2002) argue that solely controlling for prior attainment without taking into account the differences in student socioeconomic and demographic characteristics between schools could unfairly punish schools with educationally disadvantaged students and reward schools with educationally advantaged students. In this study, the models that extend VA models to include student sociodemographic characteristics are referred to as "CVA-1" models. Sometimes, the CVA-1 model is further extended to adjust for contextual or compositional factors at school and/or class levels, such as school and/or class mean prior achievement, the combined influence of student socio-economic status, the grade phase of schooling, the governance structure of schools, and the size of schools (Raudenbush & Willms, 1996; Teddlie et al., 2002; Timmermans & Thomas, 2015). This study refers to these models as "CVA-2" models.

However, there are debates about the appropriateness of including some explanatory factors such as SES in CVA models. Some argue that a student's prior attainment has a more significant impact on their academic success than their background characteristics. For example, Ballou et al. (2004) found that adjusting for student demographic characteristics has little effect on the assessment of school effects when using a comprehensive range of prior attainment measures. McCaffrey et al. (2003) suggested that in school systems with heterogeneous student populations, controlling for student socioeconomic and demographic characteristics without considering prior attainment is insufficient for identifying school effects. Moreover, some arguments are against adjusting for student background characteristics because of the worries of entrenching socioeconomic inequities and providing excuses for low-performing schools (DfE, 2010; Leckie & Goldstein, 2017; 2019). Furthermore, adjusting for contextual factors may lead to over-adjustment, as the effects of these factors might already be captured by student prior attainment (Timmermans & Thomas, 2015). In addition, adjusting for contextual factors may not be feasible for all schools, especially for smaller schools or those with limited data,

which may limit the generalizability of the value-added estimates (Tucker, 2011). It is important to note that both VA and CVA models have their strengths and limitations, and their suitability depends on the research questions, the availability and quality of data, and the study context. Therefore, it is valuable for this study to examine each model to explore the optimal model for a specific context, as the optimal model may vary depending on the datasets from different contexts.

In short, the choice of explanatory variables in value-added modelling is crucial as it directly affects the estimation of school performance measures in terms of size and significance. Therefore, researchers need to exercise caution and use appropriate statistical techniques to select the most relevant and meaningful explanatory variables for the intended purpose. This issue will be reviewed and discussed in the next section.

2.4.4.4 Differential Effects

Taking into account the range and extend of school and class effects on student outcomes, it's crucial to consider the consistency of school and class effectiveness. A sub-theme within the consistency issue is the presence of differential effects. According to Reynolds et al. (1996), one form of differential effects is the variation in the impact of schools and classes on different academic units within a school. Essentially, this examines how the size of school and class effects may differ depending on the type of subjects assessed, such as Chinese, Mathematics, and English in this study. Earlier research, including works by Mortimore et al. (1988), Thomas et al. (1997), and Thomas and Mortimore (1996), has revealed a lack of strong consistency in school and class effectiveness when different subjects are considered as academic outcomes.

Given the existence of varying school and class effects across different academic subjects, it would be meaningful for this study to extend its scope and explore the differential school and class effects across three distinct academic subjects (Chinese, Mathematics, English). Academic subjects are characterized by unique curricula and teaching methods, making it reasonable to expect that specific interventions or support may be more effective when tailored to the particular features of each subject. By providing insights into these differential effects, this study can demonstrate how VAMs can provide information for schools and classes to address the diverse needs and abilities of students across different subjects. Policymakers may also benefit from this information as it can guide resource allocation and the

formulation of policies aimed at addressing disparities in student outcomes within different academic areas. Therefore, this study may have the potential to contribute to present more comprehensive information of school and class effectiveness.

2.4.4.5 Identifying Variables Employed in Value-Added Models

One issue associated with outcome measures in the value-added models is the range of outcomes used. Researchers such as Goldstein (1993) and Thomas (1998) argued that it is important to include not only total scores in high stakes examinations, but also scores in individual subjects. In this study, student SHSEE Total score is the sum of individual student scores in three subjects (Chinese, Mathematics, English). Looking at the student outcome measure of the total score is valuable because it provides a holistic view of a student's overall performance across multiple subjects. This comprehensive assessment allows for a more complete understanding of a student's academic achievements and progress, as opposed to focusing on individual subject scores in isolation. On the other hand, there is evidence of significant departmental differences in terms of effectiveness, and using only total scores may not reveal these differences (Zhang, 2016). Moreover, it has been suggested that all prior attainment variables, including individual subject scores, should be included as explanatory variables in value-added models. This recommendation is based on the understanding that student prior attainment, encompassing performance in different subjects, is a significant predictor of student achievement (Thomas, 2001). Therefore, including both total scores and scores of several key subjects can provide comprehensive information for this study and enhance its accuracy of estimation.

To enhance the validity and accuracy of estimation in value-added models, two main concerns need to be addressed when identifying explanatory variables: the presence of a statistically significant relationship between variables considered outside the control of schools and student academic outcomes, and the differential distribution of these variables among schools. To some extent, the degree of relationship and variation between these variables is positively related to the desired effect of adjustment (OECD, 2008). The following are the several student sociodemographic characteristics typically used in value-added models.

First, several studies indicate the presence of gender disparities in academic performance across different subjects. Lynn and Mikk (2009) found that girls outperformed boys in verbal tasks, while

Hyde and Linn (2010) observed that boys tended to perform better than girls in quantitative tasks. OECD (2007) reported that girls performed better in reading and writing literacy, whereas boys performed better in mathematics and science. In primary schools, Mortimore et al. (1988) found that girls outperformed boys in most subjects, and in secondary schools, Thomas et al. (1994) reported that girls performed better in terms of the total score of GCSE. Consequently, involving gender as an explanatory variable in value-added models can provide valuable information for specific gender groups and support evidence-based policies and school improvement efforts (OECD, 2008).

Second, due to research questions, sample acquisition, and contextual factors, there lacks consensus on the appropriate utilization of various measures of student socio-economic status (SES) in the value-added models. It is widely acknowledged that family income, parental education, and occupation are commonly used to measure SES (Sirin, 2005). Some studies have also included measures of home environment indicators, such as the number of items and books in the household (OECD, 2019). Additionally, some researchers have investigated the impact of SES composition of the student on their studies (e.g., Blakey & Heath, 1992; Sammons et al., 1996). However, other studies have not found evidence of such SES effects at individual or school levels (e.g., Bondi, 1991; Hauser, 1970 ; 1971; Mortimore et al., 1988; Trower & Vincent, 1995). Regarding the contradictory results, Reynolds and Teddlie (2002) suggest that if there is little variance in SES of students in the study setting, or if researchers conduct a study in one particular SES context, then it is less likely that a compositional effect will be found. In relation to the context of this study, while characteristics such as immigrant status, ethnicity, and spoken language are commonly used in value-added models in Western countries (OECD, 2008; Leckie & Prior, 2022), such variables may not vary significantly in the context of this study. Similarly, while eligibility for free school meals is frequently used in Western countries (e.g., Ballou et al., 2004; McCaffrey et al., 2003; Thomas, 1998; Thomas & Mortimore, 1996), it is not applicable to this study. Overall, the selection of student sociodemographic variables in value-added modelling lacks standardization across countries (Bollen et al., 2001), highlighting the importance of conducting research in local contexts such as China to provide original evidence.

Furthermore, adding adjustments for class and school context variables may further isolate the contribution of schools. These variables may involve aggregations of student-level variables, such as

the mean prior attainment of all students aggregated at class and school level (De Fraine et al., 2003; OECD, 2008), or the percentage of students in a school across all prior attainment measures (e.g., Peng et al., 2006; Salim, 2011). Some variables are specific to the school or class level (e.g., community type of school, governance structure of school, class characteristics) (Mokonzi et al., 2020; Muñoz-Chereau & Thomas, 2016; OECD, 2008; Salim, 2011). However, some studies found that adjusting for these contextual effects has relatively little impact (Leckie & Prior, 2022; Marks, 2021; Timmermans et al., 2011). It should also be borne in mind that the inclusion of school context factors in the form of student composition variables presents a complex dilemma with no easy solutions. For example, Timmermans and Thomas (2015) pointed out that controlling for school context in the form of student composition may make it difficult to identify the specific areas of effective or ineffective teaching and learning within schools. On the other hand, not accounting for those school context variables may result in unfair treatment of schools with disadvantaged student populations.

In short, it should be borne in mind that the more explanatory variables are included in a model, the danger of over-adjustment increases (OECD, 2008). Moreover, the complex findings underline the importance of testing the significance of explanatory variables in developing value-added models for this study (Thomas & Mortimore, 1996).

2.4.5 Multilevel Modelling (MLM) for Data Analysis

2.4.5.1 Advantages of MLM

This study employs MLM, a powerful statistical technique commonly employed in VAMs studies (Goldstein, 1997; 2011), to estimate school and class effects on student progress. It has several advantages over traditional linear regression models.

Firstly, earlier studies examining school effects often failed to consider the hierarchical structure of schooling data (Raudenbush & Bryk, 1986). Researchers often utilized data aggregated at the school level or disaggregated school-level variables to the individual level, which may lead to potential inaccuracies in findings (Aitkin & Longford, 1986; Bryk & Raudenbush, 1992; Muijs & Brookman, 2015). MLM addresses this issue by accounting for the hierarchical nature of the data, allowing for the simultaneous analysis of individual-level variables (e.g., student background characteristics) and group-level variables (e.g., school policies). Secondly, traditional linear regression models assume

independence among observations. However, in a school setting, students within the same classroom or school may be more similar to each other than to students in other classrooms or schools. MLM can take account of the lack of independence assumption, which can lead to more accurate estimates of the school and class effects. (Goldstein, 1997; Creemers et al., 2010). Furthermore, MLM can handle missing data more effectively than traditional linear regression models. This is particularly important in education research, where missing data is common due to student mobility, absenteeism, or other factors (Muijs & Brookman, 2015).

Overall, by modelling nesting effects, MLM allows researchers to gain a better understanding of how individual-, class-, and school-level factors influence student outcomes. It can also provide more accurate estimates of predictor effects by accounting for the variation within and between groups. However, it is important to be aware of potential disadvantages and methodological issues related to MLM, which will be discussed in Chapter 3.

2.4.5.2 Equations of Multilevel Models

Equations of Raw model for three- level data structures can be described as:

$$y_{ijk} = \beta_0 + v_k + u_{jk} + e_{ijk} \qquad (1)$$

$$v_K \sim \mathrm{N}(0,\ \sigma_v^2)$$

$$u_{jK} \sim N(0,\ \sigma_u^2)$$

$$e_{ijK} \sim N(0,\ \sigma_e^2)$$

Equation (1) is a three-level variance-components model, where i = 1, …, n_{jk}; j =1, …,J_K; k= 1, …, K. n_{jk} is the number of students taught in the j_{th} class within k_{th} school, J_k is the number of classes in k_{th} school, and K is the number of schools. y_{ijk} refers to the SHSEE score of the student i taught in the j class in school k. β_0 is the overall mean of students' achievement. u_{jk} represents the effect of class j in school k on student achievement and v_k represents the effect of school k on student achievement. The variance of the school-specific and class-specific deviations (v_k and $u_{jk)}$ are of particular interest in this study. For example, Variance Partition Coefficient (VPC) measures the proportion of variance in the outcome variable that can be attributed to the different levels of a hierarchical model (Goldstein et al., 2002),which is defined as:

$$\text{VPC}= \frac{\sigma_v^2}{\sigma_v^2+\sigma_u^2+\sigma_e^2}$$

The computation of variances for each level using the raw model can serve as a reference for determining the percentage of variance explained per level. In the subsequent stages, key explanatory variables are incorporated into the model. By comparing the variances for each level in these models to the variance components in the raw model, the researcher can evaluate the extent to which the variance is accounted for in each individual level (Goldstein, 2003; Luyten & Sammons, 2010).

One of the commonly used multilevel model is the random-intercept model. It assumes that the intercept, which is the value of the dependent variable when all independent variables are set to zero, is allowed to vary across groups, while slopes remain constant across all groups. The equation of random-intercept models with covariates can be described as:

$$y_{ijk} = \beta_0 + \beta_1 X_{ijk} + v_k + u_{jk} + e_{ijk} \qquad (2)$$

$$v_K \sim \text{N}(0,\ \sigma_v^2)$$

$$u_{jK} \sim N(0,\ \sigma_u^2)$$

$$e_{ijK} \sim N(0,\ \sigma_e^2)$$

Equation (2) includes just one explanatory variable as an example, which is measured at the student level. As discussed in above sections, the analysis in this study will involves multiple variables measured at three levels. Equation (2) has a fixed part (the intercept and the coefficient of the explanatory variable times the explanatory variable) and a random part ($v_k + u_{jk} + e_{ijk}$). The random intercept means that the intercept for the overall regression line is still β_0 but for each school line the intercept is $(\beta_0 + v_k)$.

It should be noted that one of the assumptions of the random intercept model is that the effect of the explanatory variable on the dependent variable is consistent across all schools. However, there is a possible situation that the relationship between the explanatory variable and the dependent variable varies from school to school. In other words, the explanatory variable may have a large effect on the dependent variable in some schools and a small effect in others. Therefore, instead of assuming that the slope is fixed, the random-slope model allows for heterogeneity in the effect of the independent

variables on the dependent variable (Goldstein, 2003). This study employed a random intercept model rather than a random slope model for several reasons. Firstly, the main objective of this study is to generate meaningful and relevant findings that contribute to the broader research literature in China and advance understanding of the application of value-added methods in local school evaluation practices. As such, the primary focus of this study is not on the technical or methodological aspects of statistical analysis. Therefore, the study has chosen to replicate the approach used by some previous studies (e.g., Fan & Gao, 2019; Mokonzi et al., 2020; Muñoz-Chereau & Thomas, 2016). What's more, considering the limitations in scope, resources, and time inherent in EdD research, the random intercept model is simpler to specify and may facilitate more effective communication of the results of the analysis to non-technical policymakers or educational practitioners. However, employing the random slope model in future research would provide further insights.

Overall, a school with a positive value-added score (e.g., intercept residual/error term) and a 95% confidence interval above zero indicates that the school performs significantly above average, while a school with a negative value-added score (residual) and a 95% confidence interval below zero indicates that the school performs significantly below average. The confidence intervals provide information about statistical uncertainty and the inherent imprecision in calculating value-added. Hence, the information concerning the 95% confidence interval is crucial in evaluating whether the estimation result is likely to have occurred by chance. It is the foundation for assessing whether a school's performance is statistically significantly different from the typical school (Leckie & Prior, 2022; Muijs & Brookman, 2015; OECD, 2008; Thomas, 1998).

2.4.5.3 Levels Developed in MLM

As illustrated above, this study will employ three-levels models (student, class, and school level). Although some studies use two-level models (student, school level), many include more than two levels, such as the school and student levels, with the addition of the classroom level (e.g., Hill et al., 1995; Kyriakides et al., 2000; Mokonzi et al., 2020; Muijs & Reynolds, 2000; Opdenakker & Van Danme, 2000). In some cases, researchers also employ the local authority or national level as a higher order level in their studies (e.g., Munoz-Chereau & Thomas, 2016; Tymms et al., 2008). Opdenakker and Van Damme (2000) found that excluding relevant levels could lead to errors in estimating

variances across different levels. Therefore, Muijs and Brookman (2015) suggested including all relevant levels in the analysis of education data, even if no variables are collected at those levels. Kyriakides and Charalambous (2004) noted that it is essential to test the fit of various models to the data to determine the most appropriate number of levels to include.

Based on the above discussion, this study will measure both school and class effects. Including the class level has the potential to improve the fit of value-added models, and this will be tested in the quantitative analysis in Chapter 4. Moreover, previous research has found differences in school and class effects across different countries (De Fraine et al., 2003), indicating that the findings of this study can provide information to examine these differences in a new local context. It should be noted that multilevel modelling inevitably has statistical and methodological limitations, and these issues will be discussed in Chapter 3.

2.5 Challenges of VAMs

As discussed in Section 2.4.3, VAMs offer advantages over raw unadjusted attainment measures by providing a fair and more valid measure, as well as generating useful information for diagnostic and improvement purposes (Ballou et al., 2004; Goldstein, 1995; Sammons, 1999; Thomas, 1998; Thomas et al., 2007). However, VAMs are not without challenges.

One challenge is the doubt about the correct model specification. This involves identifying the appropriate variables that best fit the data (Goldstein & Leckie, 2008). Hibpshman (2004) noted that omitting causative variables could lead to biased estimates, meaning that VAMs reflect not only the school effect but also the influence of other potentially significant factors that were not controlled for in the analysis. Furthermore, VAMs, in line with raw unadjusted attainment measures, are based on past performance and provide retrospective assessments. Thus, they are helpful for institutional review and school improvement activities rather than making projections about future performance (Sammons, 2010).

In considering the practical implications for school improvement, the complexity of the statistical models may present challenges in terms of public understanding (Leckie & Prior, 2022). Simpler models are generally more transparent and might be favored by the end users, although they may be

less desirable technically (OECD, 2008; Leckie & Prior, 2022). In addition, the cost of data collection and construction as well as maintenance of is the data are also considered as implementation issues. Nevertheless, as Sammons et al. (2016) stated, the limitations of value-added methods using MLM do not necessarily mean that they should be rejected in favour of simpler approaches.

So far, this chapter has examined the development of SER, the conceptual framework employed in this study, the definition of school effectiveness, and the emergence and application of VAMs to evaluate school effectiveness. The discussion has primarily drawn on international literature or literature from mainstream SER countries. However, as discussed in the previous chapter, the field of SER/EER has become increasingly diverse across international contexts. Given that this study is conducted in the local context of China, where SER/EER is under-presented, it is essential to review relevant research conducted in mainland China to inform the design of this study.

2.6 Value-Added Evaluation Research in Mainland China

2.6.1 Brief Review of the Development of Value-Added Evaluation Research

Despite the trend of SER/EER toward internationalization, there have been relative few empirical studies on VAMs conducted in mainland China before 1990. Chinese scholars began paying attention to VAMs in the 1990s, with some scholars translating relevant articles from Western countries into Chinese (e.g., Zhan, 2001; Zhang, 1998), and others conducting empirical studies in collaboration with researchers from other countries (e.g., Peng et al., 2006). However, the development of VAMs research in China experienced slow growth, possibly due to limitations in statistical analysis techniques, a shortage of human resources with data handling skills, and relatively less attention from the government (Guo & Wang, 2021).

Since 2005, there has been a gradual increase in the number of empirical studies focusing on VAMs in China. Researchers have paid attention to methodological techniques used in international VAMs research and some have employed multilevel modelling in relevant Chinese literature (e.g., Thomas & Peng, 2011; Thomas et al., 2012; Thomas et al., 2015; Yang & Zhang, 2022). Although some studies have focused on senior high schools and postgraduate institutions, empirical research on junior high schooling remains relatively scarce (Bian & Lin, 2007). The introduction of the Overall Plan on

Educational Evaluation Reform in 2020 (State Council, 2020) has led to an increase in the number of studies related to value-added evaluation in mainland China. Reviewing the literature shows that researchers are paying more attention to the value and implementation of VAMs in practice (e.g., Chen & Dong, 2022; Li et al., 2022; Zhang et al., 2022). However, there is still a lack of longitudinal empirical studies specifically examining the implementation of VAMs (Guo & Wang, 2021).

In short, the review of the development of VAMs research in China reveals the limited empirical research that explores VAMs focusing on junior high schooling and the lack of empirical research that explores the implementation of VAMs. The following sections will review typical literature associated with investigating VAMs as a fairer school evaluation method and the considerations of implementing VAMs in the context of China.

2.6.2 A Brief Review of Quantitative Research in the Context of China

This section provides a brief overview of value-added evaluation studies that mainly utilized quantitative research methods within the context of China. Through the review of the typical literature, this section seeks to identify gaps in the current research and, subsequently, facilitate the design of this study. Therefore, the following table indicates the key features of several value-added evaluation studies in mainland China.

Table 1

Key features of empirical value-added evaluation studies in mainland China (Sc = schools; St= students)

Study	Phase of schooling	Region	Sample size	Some Explanatory variables	Level of models	Outcome measures
Fan & Gao (2019)	SHS	East	Sc-43 St-3193	school entrance exam scores; gender; student origin (rural/urban); SES; school type	Two	Chinese, Mathematics, English
Liu (2022)	JHS	East	Sc-39 St-3909	gender; SES; location of household registration	Two	Mathematics
Lv (2015)	JHS	West	Sc-21 St-6445	gender; school size; school type	Two	Chinese, Mathematics, English
Peng et al. (2006)	SHS	North	Sc-17 St-9247	school entrance exam scores; gender; student type; student origin (rural/urban); whether a student's total achievement score was in the top 25%	Two	Total, Mathematics, English
Peng et al. (2013)	JHS	North	Sc-25 St-4053	school entrance exam scores; gender; SES; school mean student prior attainment and mean SES	Two	Chinese, Mathematics, English
Shao et al. (2021)	JHS	North	Sc-33 St-4336	school entrance exam scores; parental occupation and education qualification; items in households	Two	Chinese, Mathematics, English
Thomas et al. (2012)	SHS	West and East	Sc-120 St-90000	school entrance exam scores; gender; age; SES; school type	Two	Total, Chinese, Mathematics, English
Thomas (2020)	SHS	West and East	Sc-120 St-90000	school entrance exam scores; gender; age; SES; school type	Three	Total, Chinese, Mathematics, English
Xin et al. (2012)	SHS	North	Sc-40 St-2132	school entrance exam scores; gender; student origin (rural/urban); school type; school size	Two	Chinese, Mathematics, English
Wang et al. (2009)	SHS	West	Sc-30 St-16552	school entrance exam scores; gender; learning capacity; school type; school size	Two	Chinese, Mathematics, English

Based on the presented table, it can be observed that several empirical studies on VAMs have been conducted in the North, East, and West regions of China. However, there is a notable lack of research in the South, which is the area where this study is being conducted. Out of the 10 studies reviewed, 4 focus on junior high schooling, indicating the need for more research evidence concerning the evaluation of junior high school performance. This aligns with the need for reform in compulsory education evaluation as mentioned above (junior high schooling falls within the compulsory education stage in China).

The findings of the studies reported in Table 1 reveal significant and substantial differences in the estimated value-added effectiveness of junior high schools and senior high schools. These differences vary not only across regions but also across subject outcomes. For instance, some studies (e.g., Fan & Gao, 2019; Xin et al., 2012) found that up to 35% of the overall variance in student raw attainments across three subjects could be attributed to differences between schools. However, other studies (e.g., Peng et al., 2006; Peng et al., 2013; Shao et al., 2021) reported findings similar to those of Thomas (2020), where differences between schools accounted for up to 27% of the variance. Notably, Thomas et al. (2012) and Thomas (2020) had much larger sample sizes compared to other studies and employed three-level multilevel models rather than the two-level multilevel models used in other studies. Thus, they appear to provide more rigorous estimates of the size of school effects and may serve as a valuable reference for this study. Furthermore, Thomas (2020) suggests considering the complete complexity of educational effectiveness when measuring the school performance and provides evidence that employing a CVA model may reduce the size of school and class effects. This informs this study to enhance its research evidence by considering the methodology in the following ways.

Regarding methodology, most of the studies reported in Table 1 use two-level multilevel models, with the exception of Thomas (2020) under the IEEQC project, which developed three-level models. However, the IEEQC project mainly focused on SHS. It highlights the importance of developing three-level multilevel models and examining JHS as demonstrated in the previous section (Section 2.4.5.3). Of 4 studies examining JHS, only 2 employ student prior attainment data. As discussed in Section 2.4.4.2 about the importance of adjusting for student prior attainment, it underscores the value of this study in employing longitudinal data over a three-year period. In terms of explanatory variables,

student prior attainment, gender, and socioeconomic status are typically included in the models. Parental education qualification and parental occupation are also commonly collected in most studies. In contrast to VAMs studies in Western countries, student background characteristics such as ethnicity, mobility, and eligibility for free school meals are not commonly used in Chinese literature (e.g., Goldstein et al., 2008; OECD, 2008; Thomas, 1998; Thomas & Mortimore, 1996).

In terms of outcome measures, Chinese, Mathematics, and English subject scores are the most employed measures in these studies. However, total score outcomes are less frequently used. In line with the discussion on using total and different subject outcome measures (Thomas & Mortimore, 1996), this study should consider using both total scores and scores for Chinese, Mathematics, and English as outcome measures. Furthermore, it is important to note that Yang and Zhang (2022) have highlighted a scarcity of studies demonstrating the validity of different fitted value-added models used in research. This scarcity is also found from above studies. Therefore, it is crucial for this study to contrast the findings of different value-added models to illustrate the development of those models and their ability to accurately fit the data.

2.6.3 A Brief Review of Studies Concerning the Implementation of VAMs

It is evident from the limited studies available in the Chinese context that there is a consensus among researchers regarding the advantages of implementing VAMs in China (Guo & Wang, 2021). For example, value-added approaches allow for schools to be compared more fairly (Ma & Peng, 2006), while providing fair and accurate evaluation information that can help improve the allocation of educational resources (Zhen & Song, 2021). Du and Hao (2021) note that VAMs can provide comprehensive information about differential effects for different groups of students or different departments. Xin and Li (2020) contend that the main advantage of implementing value-added evaluation is to provide valuable diagnostic information about education and teaching practices, which can help educators become more active in promoting the development of students, rather than being used for determining rewards or penalties for schools or teachers.

However, alongside the advantages, researchers have also discussed the disadvantages of VAMs and expressed concerns about its implementation. One common concern is the unintended side effects of "teaching to the test," where the pressure to achieve good test scores can lead to a narrow focus on

exam preparation and neglect of broader educational goals (e.g., Ma, 2020; Zhen & Song, 2021). Scholars argue for the inclusion of non-academic subjects in student outcome measures to align with the government's emphasis on students' all-around development. This view is shared by Xie and Zhang (2021), Gao and Song (2021), and Peng and Zhang (2021). Other concerns expressed by many researchers include ethical issues related to data collection, increasing burden on schools and teachers, and the lack of appropriate training to interpret and effectively use evaluation results (e.g., Feng & Zhou, 2022; Ren, 2022). These issues highlight the potential challenges and limitations that need to be considered when implementing VAMs in the Chinese educational context.

While there is increasing discussion in academic circles about implementing VAMs, the views of important stakeholders such as policymakers, headteachers, and teachers are rarely presented in literature. Since these stakeholders play a critical role in implementing VAMs (OECD, 2008), it would be valuable for this study to collect their views (specifically policymakers and headteachers) to provide new qualitative research evidence that can inform policy decisions.

In short, a considerable proportion of VAMs studies in China are exclusively quantitative in nature. These studies tend to concentrate on identifying and estimating school effectiveness within specific phases or regions of research (Feng & Zhou, 2022; Guo & Wang, 2021). Moreover, discussions on implementing VAMs are mostly limited to scholars' perceptions, while responses from practical stakeholders are rarely presented. Therefore, to extend the literature on VAMs in China, it is meaningful to conduct empirical research to test the potential of implementing VAMs in a new regional context, given the vast educational landscape in China. Chapter 3 will outline the research design, which is informed by the literature review conducted in this chapter. The last section of this chapter will provide contextual information about the study.

2.7 Overview of the School Evaluation Context of China

2.7.1 Key Stages of Schooling

The general structure of the Chinese education system is very much in line with prevailing international practice. Preschool, or kindergarten, can last up to three years, with children entering as early as age 3 and staying until age 6, when they typically enter primary school. According to *Compulsory Education*

Law of the People's Republic of China, there are nine years of compulsory education in China, typically for ages 6 to 14. Students transfer to a different school during mandatory schooling at age 11. Specifically, primary education covers students aged 7 to 11, encompassing the first through sixth grades. Following primary education, students aged 12 to 14 engage in junior high education, comprising the seventh through ninth grades, which is the specific focus of the present study. There are two streams of education after completing nine years of compulsory schooling. One is general academic, and the other is vocational. The academic track is an additional three years of non-compulsory education offered in senior high schools from age 15 to 18. Usually, students who are successful in the university entrance examination attend university for four years, while students who fail to enter university may join a post-secondary institution for a three-year certificate or diploma course. On the other hand, vocational education is provided by four types of vocational school, regular specialized schools, adult specialized schools, vocational secondary schools, and crafts schools.

2.7.2 Measuring Educational Outcomes Through JHSEE and SHSEE

China has been successful in ensuring that children attend school during the compulsory education stage, which runs from Grade 1 to Grade 9 (ages 6/7-14/15) (Liu & Teddlie, 2003). Although senior high school education is not compulsory, Chinese families often prioritize supporting their children in pursuing further education. As a result, the Senior High School Entrance Examination (SHSEE) plays a crucial role in determining admission to senior high schools in the first or upper average tier. The SHSEE is a unified examination taken by students in a particular city at the end of their junior high education, typically in Grade 9 (age 14/15). This exam is similar to the Junior High School Entrance Examination (JHSEE), which is taken in Grade 6 (age 11/12).

2.7.3 Education Governance in Chinese Education System

In China, the education system uses a regulatory approach to ensure school quality (OECD, 2020). The Ministry of Education has established school management standards for compulsory education (COME, 2017), which serve as a basis for regional educational authorities to develop local regulations. However, these national standards are often broad and lack specific measurable indicators, making it challenging for regional authorities to create effective regulations (Xin, 2016). While performance-based public accountability, which relies on student outcomes such as test scores to hold schools

accountable, is commonly used in other education systems, it is not prevalent in China (OECD, 2020). Instead, the Chinese education system initiates a national assessment scheme (National plan for monitoring compulsory education quality) (State Council, 2015) to internally monitor its national education quality at the basic education level.

School evaluations in China also rely heavily on internal school inspections (evaluation results are not available publicly), which are mandated by the central government and carried out by education inspection committees (OECD, 2020). These committees operate at different administrative levels, from the central government down to the local levels (the province, the city, the district) and are responsible for improving the quality of education in China. The authority for school resources, curriculum provision, and student admissions lies primarily with local and regional authorities (at province or city level), rather than with school principals, teachers, or governing boards. Consequently, school autonomy in China is relatively low compared to OECD countries, with headteachers having limited involvement in policy formulation, resource allocation, and course offerings (Gao et al., 2006; OECD, 2020; Zhen, 2019). Despite these differences in governance, the aim of enhancing education quality through school inspections remains paramount.

Overall, the use of unadjusted raw high-stakes examinations scores to assess school effectiveness internally has been prevalent for a long time in China (Xin, 2016). However, with education reforms and a growing focus on students' all-around development, new evaluation approaches, such as the Student Growth Record Report, 'Green Indicators' model, and Comprehensive Evaluation Framework, have emerged to support school improvement (Liang et al., 2016; Xin, 2019). For example, Green Indicators were used in the Shanghai education authority to inspect and evaluate school annually. The main objective of this administration system was to assess school quality holistically (Liang et al., 2016). As Bian and Lin (2007) claimed, the value of the green idea in school evaluation is to emphasize the progress of students made in schools and schools need a comprehensive and sustainable development. Student Growth Record Report is based on educational and teaching objectives and consciously collects works and other evidence related to student performance in various subjects. Through reasonable analysis, it reflects the strengths and weaknesses of students in the learning and development process, reflects the efforts and progress made by students in achieving their goals, and

through student reflection and improvement, motivates students to achieve higher achievements (Ma, 2021). Compared to traditional result-oriented evaluation methods, these evaluation approaches focus more on the development of students and provide a process-oriented evaluation of students or schools. The data collected under this concept and evaluation methods can provide potential conceptual understanding and data foundation for the development of value-added evaluation methods.

Overall, while VAMs are a widely accepted and reliable evaluation method increasingly used in education systems worldwide, they have not been generally developed to measure school effectiveness in China. This context has raised interest of this study to explore the feasibility of applying VAMs in China's context.

2.8 Chapter Conclusion

The review of international literature highlighted the variations in school and class effects across countries, emphasizing the need for new original evidence to explore the generalizability of VAMs and potential unique findings relevant to the specific context (Lindorff et al., 2020; Thomas, 2001). Previous research also demonstrated the importance of providing detailed findings generated by various value-added models to inform policy decisions (Leckie & Goldstein, 2019; Leckie & Prior, 2022; Marks, 2017; 2021; Muñoz-Chereau & Thomas, 2016; Thomas & Mortimore, 1996; Thomas, 2001; Timmermans et al., 2011). However, research on this topic in China is scarce, emphasizing the significance of this study.

Regarding the research methodology, the literature review highlights the importance of combining quantitative and qualitative analyses to understand the potential benefits of implementing VAMs and improving school evaluation practices. Although studies from Western countries have explored the perspectives of policymakers and practitioners (e.g., Muñoz-Chereau et al., 2020; Peng et al., 2013; Sammons et al., 1998; Saunders, 2000), there is a scarcity of research on practitioners' views in China. Therefore, obtaining qualitative evidence from policymakers and practitioners becomes particularly valuable for understanding the potential implementation of VAMs in the local context.

Overall, this literature review underscores the importance of generating new empirical evidence through quantitative and qualitative approaches, thereby seeking to bridge the gap between research

and practice and support school evaluation and improvement. This chapter has established the foundation for the current study by reviewing relevant literature, theories, policies, and practices in the Chinese context. The next chapter will build upon this groundwork to outline the study's philosophical underpinnings, research design, and methods for data collection and analysis.

Chapter 3 Research Design and Methodology

3.1 Chapter Introduction

This chapter presents the methodology and methods used to address the study aim and research questions. It begins with a justification of the philosophical standpoint used for this research. Subsequently, the sequential mixed methods research design that has been employed in this study is presented in detail. It then proceeds with a description of the sampling process, followed by an explanation of the two data collection methods and the data analysis techniques employed. Finally, the chapter concludes by discussing ethical issues and potential limitations of the methodology.

3.2 Research Questions

RQ1: What is the range and extent of school and class academic performance in 11 junior high schools in the W district in Southwest China based on Raw, Value-added (controlling for prior attainment only), Contextual value-added (controlling for not only prior attainment but also student background characteristics, class, and school context) measures?

RQ2: What are stakeholders' perceptions of the potential advantages, disadvantages, and implementation of the value-added approach in school evaluation practices when local contexts are considered?

3.3 Philosophical Standpoint

The philosophical standpoint adopted in this study is pragmatism, which allows for flexibility in utilizing different paradigm standpoints (Tashakkori & Teddlie, 1998). Historically, most school effectiveness research was situated within a quantitative-oriented or primarily quantitative tradition, which is based primarily on a positivist worldview (Chapman et al., 2015; Creemers et al., 2010). Within the positivist standpoint, reality is considered objective and can be understood by examining the laws by which it is governed. Therefore, research within this paradigm typically aims to test theories, directly observe and quantitatively measure phenomena, and objectively predict relationships between variables to produce universal knowledge through statistical generalization (Tashakkori & Teddlie, 1998). Constructivism, which is one of the foundational philosophical standpoints guiding

qualitative research, views that the world and knowledge as shaped by social and contextual understanding. Instead of generalization, constructivists argue that there are multiple unique realities and many different phenomena, and relationships are viewed as social constructs by which an individual makes sense of the external world or reality. Thus, research within this philosophical paradigm typically aims to answer questions about 'why' or 'how' (Creswell & Clark, 2011; Hadfield & Chapman, 2015).

Pragmatism emphasizes addressing practical problems and using methods that work to find solutions. It does not take a definitive stance in favor of positivism or constructivism (Creswell, 2003). Therefore, the researcher may choose methods from any paradigm that is suitable for answering the research questions (Johnson & Onwueghbuzie, 2004). Tashakkori and Teddlie (1998) connect pragmatism to mixed-methods research, highlighting its flexibility in investigating multiple research questions and allowing researchers to employ both quantitative and qualitative methods in a single study.

As indicated earlier, pragmatism supports utilizing methods that are most appropriate for addressing the study's objectives and research questions. Therefore, adopting pragmatism in this study yields two key implications. Firstly, to investigate the reliability and validity of value-added approaches in measuring school academic performance in China, quantitative methods and value-added models are employed to estimate school and class effects on student outcomes. Secondly, qualitative methods are employed to focus on the perspectives of stakeholders and explore the integration of VAMs into school performance evaluation practices.

3.4 Research Design

The pragmatist paradigm is often regarded as the most suitable paradigm for mixed methods designs, particularly when the research world view encompasses multiple world views (Creswell & Clark, 2011). Mixed methods research combines procedures typically employed in both quantitative and qualitative studies to achieve the broader goals of gaining deep understanding and corroboration (Johnson et al., 2007). For this study, a mixed methods approach was adopted to leverage the synergies and strengths of both quantitative and qualitative methods and gain a more comprehensive understanding of value-added approaches than would be possible with either method alone (Teddlie & Sammons, 2010).

This study employed a sequential mixed methods research design. The first phase focused on answering RQ1 through quantitative methods, which involved data collection and statistical analysis. Datasets, such as student prior attainment, academic outcomes, background variables, as well as class and school contextual variables, were used in value-added models to estimate the effects of schools and class on student achievement (e.g., Kyriakides & Creemers, 2008; Teddlie & Sammons, 2010).

In the second phase, qualitative interviews were conducted with policymakers and headteachers to address RQ2. As noted by Teddlie and Sammons (2010), if the inclusion of qualitative methods in a study can provide additional value, it is a good rationale for using mixed methods in the study. In this study, there were three main reasons for collecting qualitative data. Firstly, collecting qualitative data from stakeholders was valuable in exploring the specific local context of China and the factors that could either facilitate or hinder the implementation of value-added approaches in school evaluation practices. Moreover, qualitative findings are likely to be helpful in guiding practitioners by providing insights into practical decision-making processes for improving school evaluation methods within a given local context (Hadfield & Chapman, 2015). It is worth mentioning that the research design employed in this study is not without its limitations and ethical considerations, which are discussed in Sections 3.7 and 3.8. To obtain both quantitative and qualitative data, research instruments were developed, and the sampling approaches for the two phases is explained separately in the following sections.

3.5 Development of Research Instruments

3.5.1 Quantitative Data and Student Questionnaire

The process of estimating the impact of a school or class on student progress is a complex undertaking. In line with the theoretical conceptual framework guiding the value-added research design, this study collected data from multiple sources, including at student, class, and school levels. The datasets comprised individual student outcomes as well as explanatory variables derived from a thorough review of international and local research on school and educational effectiveness, as detailed in Chapter 2. Notable sources include Chapman et al. (2015), OECD (2008), Peng et al. (2006), Reynolds and Teddlie (2002), Scheerens et al. (2003), Thomas (1998; 2015; 2020), and Thomas and Mortimore (1996).

To conduct value-added measures, it is necessary to have both baseline and outcome attainment data over a specific period (Scheerens et al., 2003). As previous studies have indicated (Goldstein, 1993; Thomas, 1998), student attainment in standardized and externally marked examinations is suitable for VAMs. Accordingly, this study used the prior attainment results of students' JHSEE examinations (taken by all students in the district in 2018 at the end of Grade 6, aged 11/12) in Chinese, Mathematics, and English as the baseline attainment. For outcome attainment, the study used students' SHSEE examination results (taken by the same cohort of students three years later in 2021 at the end of Grade 9, aged 14/15) in the same subjects. All raw scores were transformed and calculated using a standardised approach (mean=0, standard deviation=1) and attainment data was matched over time using unique individual ID codes. It is worth noting that including total examination score and other curriculum subject scores is highly recommended to account for potential differences in departmental effectiveness. If only total scores for a school were used, it would be difficult to detect school effects on student outcomes in specific subjects (Goldstein, 1993; Thomas & Mortimore, 1996).

In addition to the prior attainment data, this study included additional explanatory variables that encompassed background information about individual students (e.g., gender, SES), as well as contextual information about classes and schools. These variables were necessary for statistical controls because they represent factors outside the school's control but have a significant impact on a student's academic outcomes (Mokonzi et al., 2020; Munoz-Chereau & Thomas, 2016; OECD, 2008; Sammons et al., 1996; Scheerens et al., 2003; Thomas, 1998; Thomas & Mortimore, 1996).

The online student questionnaire was designed to collect students' demographic and family background information. In addition to its use in developing value-added models, the data collected by the questionnaire can provide systematic numerical descriptions of the sample profile of students and their socioeconomic background. The decision to use an online delivery method was motivated by its time-saving and cost-effective advantages. Furthermore, due to the influence of COVID, the LEA recommended this method for health and safety reasons.

The questionnaire consisted of two parts and 21 questions. Part 1 requested general information about the student, including their current school ID (assigned by the researcher), current class number, student ID, and senior high school entrance examination (SHSEE) registration number. This

information was necessary for accurately matching prior and outcome attainment records during data processing. Part 2 collected information about students' demographics and family background, such as gender, household registration status, possessions in the student's home, parents' education level, and parents' occupations (see Appendix 2 for the student questionnaire instrument).

Data on student attainments and contextual factors at the class and school levels were obtained from the LEA. The contextual factors included information on school composition in terms of school prior mean attainment, and class composition in terms of class mean prior attainment. Table 2 displays the variables and data collection instruments used in this study.

Table 2

Variables and data collection instruments

W district (2018-2021) 11 Schools; 46 Classes; 1,596 students	
Outcomes: SHSEE scores (total score, Chinese, Mathematics, English) Prior attainment: JHSEE scores (total score, Chinese, Mathematics, English)	Obtained from the LEA
Student background: gender, place of family household registration, number of children in the family, whether receiving private academic tuition, items in the house, items students own, education and occupation of parents	Collected by Student Questionnaire
Context: class composition information in terms of class mean prior attainment, school composition in terms of school mean prior attainment	Obtained from the LEA

Although potential issues related to data quality that may influence the validity of study findings will be discussed in section 3.7 and 3.8, there is some information worth to be mentioned here. First, the data structure of the sample described above is strictly hierarchical with three levels: students at level one, classes at level two, and schools at level three. Second, each student was assigned a unique student ID, which was used in both the JHSEE and SHSEE assessments, as well as in the student questionnaire. The unavailability of this unique ID could result in ethical concerns, which are addressed in section 3.8. Third, although there are potential concerns related to student mobility, within the context of China, the likelihood of students changing classes during the three years of junior high schooling is relatively low (Wang, 2018).

3.5.2 Qualitative Data and Semi-structured Interviews

3.5.2.1 Rationale for Using Semi-structured Interviews

To collect qualitative data, face-to-face semi-structured interviews were conducted with headteachers and policymakers (see Appendix 4 for interview question guide). The aim of the qualitative phase was to gain a deeper and more subjective understanding of how policymakers and headteachers perceive the potential benefits and disadvantages of applying value-added school evaluation methods in practice. Although interviews are a time-consuming method of data collection, their use was crucial in this study for several reasons. Firstly, interviews allow for exploration and probing of participants' thoughts, leading to more in-depth information. Participants may have varying levels of knowledge about the implications of VAMs in practice, based on their position and work experience. Secondly, interviews enable the researcher to encourage that respondents answer all questions and to probe for more information when responses are vague. This type of interview was preferred in this study as it facilitated a conversational and situational tone (Rubin & Babbie, 2017).

3.5.2.2 Semi-structured Interviews

Hadfield and Chapman (2015) have demonstrated a holistic approach to qualitative research on educational effectiveness and improvement, which informed the design of the interview schedule in this study. According to their framework, three areas can generate valuable data: practitioners' current views on their practice, their perspectives on what could be changed in the current situation, helping practitioners to contextualize new knowledge and bring about effective change. These three areas guided the structure for the interview questions. Specific interview questions were derived from the literature. The resulting interview schedule is detailed below:

Table 3

Interview schedule for policymakers and headteachers

Section	Sub-RQs of RQ2	Guiding framework	References
1	What are the views of stakeholders on the purpose of school evaluation, advantages, and disadvantages of current raw attainment measures in junior high school performance?	Practitioners' existing views of their practice.	Bee (1973); OECD (2009)

2	What are the perceptions of stakeholders on the concept of value-added and value-added measures?	Practitioners' views about what in the current situation could be changed.	Saunders (2001) Stoll et al. (2006) OECD (2009) OECD (2020)
3	What are the perceptions of stakeholders on motivations to implement the value-added approach in school performance evaluation?		Feng and Zhou (2022)
4	What are the perceptions of stakeholders on the factors that may help or hinder the application of the value-added approach?	Help practitioners contextualize new knowledge and bring about effective change.	Bee (1973) Saunders (2001) Stoll et al. (2006) OECD (2009) OECD (2013) OECD (2020) Munoz-Chereau et al. (2020)

3.5.3 Piloting

3.5.3.1 Pilot Testing of Student Survey Questionnaire

Piloting a study is crucial for testing the feasibility of the research design and identifying critical aspects of research instruments. In this study, a pilot test was conducted to ensure accessibility to participants, revise questions and answer choices, identify potential problems, and estimate the time needed for completion (Aldridge & Levine, 2001). The pilot test took place in March 2020 during term-time, and an online questionnaire was administered to 76 Grade 9 students from two public junior high schools in G City by a form tutor. The questionnaire was translated into Chinese to enable students to respond appropriately. It was piloted in two stages. In the first stage, one class of students completed the questionnaire, which was then reviewed and revised based on the feedback.

Overall, feedback from the pilot test indicated that the questionnaire structure did not require any changes. However, the item that required students to fill in the name of their school was changed, as it was found that student responses to this question were quite varied, even among students from the same class and school. To avoid any potential data errors and impact on the findings, this item was revised to allow students to choose their school's name from a list of options. Moreover, the options for parental occupations were also revised, as some students were confused about the classification of

occupations, resulting in many selecting the "other" response option. The researcher therefore revised the original classification to align with the Occupational Classification of the People's Republic of China (Ministry of Human Resources and Social Security of the People's Republic of China, 2015). The revised questionnaire was then administered to the other class of students in the second stage and reviewed and revised accordingly.

3.5.3.2 Pilot Testing of Semi-structured Interview

During the piloting phase of the interview schedule, the researcher tested it with one policymaker and one junior high school headteacher from another district in the city. The feedback from the piloting revealed that some questions needed to be revised to encourage interviewees to provide further relevant information. For example, the question "what policies and practices of junior high school evaluation in the district are currently practiced?" needed to be supplemented with questions that asked the interviewees about their views on the strengths and weaknesses of these policies and practices. What's more, it was found that it would be better to ask interviewees to provide their general perceptions first before asking specific questions related to the context of the district. For instance, the question "what is your perception of the concept of evaluation in the context of the district?" was modified to ask about interviewees' views on the purpose of school evaluation.

It was also discovered that the policymaker in the pilot was more familiar with school evaluation and VAMs, while the headteacher was less familiar with VAMs and paid less attention to the reform of evaluation methods. Thus, the researcher allocated more time to school practice-related questions with headteachers. This included asking about their views on the burden of schools under school evaluation and the availability of the school's human resources to conduct value-added evaluation. As a result, headteachers were able to provide more detailed information regarding implementing VAMs at the school level.

3.6 Data Sampling

3.6.1 Samples in the Quantitative Phase

This study aimed to utilize a complete cohort of Grade 9 students from 11 public junior high schools in W district of G city who completed the SHSEE in the 2020-2021 academic year at the age of 14/15.

Private school students were excluded from this research as their JHSEE results are not available from the LEA. All 11 public junior high schools within the district and 46 classes within those schools were focused on. While the sample size is considered small for conducting multilevel modelling value-added analysis, it is acceptable due to the inclusion of class and school level data (see further limitation discussion in page 131). It is worth noting that JHSEE and SHSEE are compulsory exams for all junior high school students, and while some students may not attend senior high school, SHSEE is also a graduation examination for junior high school students. Hence, the likelihood of incomplete student outcome attainment is low. As for the issue of student mobility across classes during junior high school, students usually remain in the class they are assigned to at the beginning of school for the three-year duration of junior high school.

The focus of this study was on public junior high schools, as they fall within the compulsory education phase and receive significant attention from the government (Ren, 2022). Moreover, this phase is closely associated with the current national education evaluation reform, which targets Grades 1 to 9 (ages 6 to 15). Given the limited scope of an EdD research project, it was deemed reasonable to focus on a single LEA in China. It can be challenging to obtain student attainment data at the city or provincial level, and existing studies in China (e.g., Thomas et al., 2012) suggest that conducting school effective evaluation research in a single LEA can be more appropriate than a national study. The selection of W district for this study was based on its representative characteristics within the city. W district was chosen as it reflects the average level of junior high schools across the entire city, without demonstrating exceptional performance or significant underperformance. By selecting a district with typical features, the study aims to provide insights that are applicable to a broader context and enhance the generalizability of the findings. Additionally, the absence of a widespread phenomenon of students moving across districts from elementary to junior high school contributes to the stability and consistency of the student population within W district, which is beneficial for the analysis and interpretation of the data. This selection acknowledges that the study sample is relatively small; however, it is argued the approach taken is adequate to address the stated aim which is essentially exploratory and that more comprehensive research results can be provided to policymakers. On top of that, the researcher is familiar with the context of W district and received the most support from the LEA to access the research data.

3.6.2 Impact of Missing Data

The potential participants are approximately 2,006 students. The working sample for this research represents a comprehensive 100% coverage of the public junior high schools in the target district, encompassing all classes within these schools, with no missing data at the school and class levels. However, when it comes to the student background characteristics collected by student online questionnaires, there is an 87% completion rate, with 1,745 students having successfully completed the questionnaire, out of which 1,596 responses were deemed valid. This means that the sample represents about 80% of the student background characteristics.

Given the presence of missing data, it is essential to examine factors that can potentially assure that the impact of these missing values on data quality is limited. First, it's important to note that the amount of missing data in this study is consistent with the levels observed in other VAMs research. For instance, a previous study conducted by Thomas and Mortimore (1996) reported a 28% rate of missing data due to incomplete prior attainment. Similarly, Munoz-Chereau (2013) found a 30% rate of missing data at the student level, and Zhang (2016) reported an average non-response rate of 34.3% across various schools for variables related to student background characteristics. Second, this study achieved 87% response rate with 90% of the collected responses deemed valid. This indicates a certain strong level of participation in the study (Rubin & Babbie, 2017). The remaining 13% non-response rate may be attributed to the voluntary nature of student participation and potential limitations in facilities or time constraints. As for the invalid responses, various factors may contribute to their occurrence, including misunderstandings of questions, lack of engagement, or technical issues. For instance, while variables like student gender and binary questions were fully observed, some students provided extreme outliers in responses related to variables such as parents' education, parents' occupation, and household items. Third, it's worth noting that the rate of missing data across the 11 schools ranged from approximately 12% to 26%, with an average of 19%. This missing data rate showed potentially consistency across the different schools, suggesting that the missing data may not stem from specific biases in data collection.

While conducting a Missing at Random (MAR) test is a valuable approach for addressing missing data, practical constraints, including time, resource limitations, and the overarching objectives of this study,

led to the decision not to perform such a test. However, it's important to acknowledge that, in the absence of a MAR test, there is the possibility that missing data could introduce bias into the study's findings and impose limitations on the generalizability of the results.

3.6.3 Samples in the Qualitative Phase

Six stakeholders, including three policymakers and three school headteachers, were voluntarily interviewed to address RQ 2 and its sub-questions. The small size of samples was used and expected to provide in-depth data for answering RQs (Creswell & Poth, 2016). Purposive sampling was employed to select policymakers and headteachers because it is believed to be more efficient than random sampling in qualitative studies (Vasileiou et al., 2018).

As discussed in Chapter 2, VAMs can improve school evaluation practices and promote school improvement if properly implemented. Policymakers and school headteachers play a critical role in promoting the implementation of VAMs, making them ideal choices for providing insights into their use. To protect the anonymity of participants, acronyms have been used in their description. Specifically, the first key participant (referred to as PM1) was chosen due to their involvement in education policy formulation, program development, the establishment of code of conduct, and the regulation of district affairs. The second participant (PM2) is responsible for overseeing educational quality, evaluating educational initiatives within the district, and providing guidance on school evaluation processes. The third participant (PM3) plays a critical role in school and teacher evaluation practices, as well as conducting research programs for professional development, school evaluation, and improvement. Given their experiences and positions, it was expected that they would provide valuable insights into the conditions and issues surrounding the application of VAMs in school evaluation in the district.

In addition to exploring policymakers' perceptions, it was considered necessary to collect qualitative data from headteachers. Headteachers are the main users of value-added data for the purpose of school improvement and are also the objects of VAMs (Saunders, 2000). Therefore, 3 public junior high school headteachers were selected from a sample of 11 schools included in this study. Among them, two headteachers are from urban schools (HT1 and HT2), while one is from a rural school (HT3) within the target district. The selection process relied on a voluntary basis, where headteachers who

expressed their willingness to participate were included in the study. The LEA played a crucial role in facilitating the research by providing the necessary contact information of junior high school headteachers, enabling the researcher to establish communication and arrange interviews. Table 4 will provide information about all six interviewees.

Table 4

Summary of the information about six interviewees

Interviewee	Acronyms	Description
Policy maker	PM1	Involvement in education policy formulation, program development, the establishment of code of conduct, and the regulation of district affairs
Policy maker	PM2	Responsible for overseeing educational quality, evaluating educational initiatives within the district, and providing guidance on school evaluation processes
Policy maker	PM3	Plays a critical role in school and teacher evaluation practices, as well as the coordination of research programs related to professional development, school evaluation, and improvement
Headteacher	HT1	From public junior high schools in urban area of the target district
Headteacher	HT2	From public junior high schools in urban area of the target district
Headteacher	HT3	From public junior high schools in rural area of the target district

3.7 Data Analysis

3.7.1 Multilevel Modelling for Quantitative Data Analysis

As previously discussed, MLM is widely recognized as an accurate and flexible tool that has been applied in a growing number of school effectiveness studies internationally to provide value-added measures of school effectiveness (Gray et al., 1990; Goldstein, 2003; Jesson, 1996; Luyten & Sammons, 2010; Nuttall, 1991; Thomas, 1998; Thomas et al., 1997). Therefore, this study employed MLM as the method of analysis. Chapter 2 reviewed the literature and highlighted that a range of multilevel statistical models (value-added models) are used to calculate school and class effects, and the results vary depending on the specific model being applied. This variation has been observed in many studies (Leckie & Prior, 2022; Munoz-Chereau & Thomas, 2016; Thomas, 2001; Thomas & Mortimore, 1996). To address this issue, instead of using a single final complex value-added model,

this study developed and discusses the progression from a simple model to a complex model. By comparing school and class effects calculated by various value-added models, the whole modelling process aimed to develop a reliable and refined model.

The models were developed sequentially, following a strategy adapted from previous research (Leckie & Prior, 2022; Mokonzi et al., 2020; Thomas & Mortimore, 1996) to assess the impact of the addition of multiple levels and explanatory variables. In the first stage, a Raw model with no explanatory variables was employed as a basic model for four outcome measures. A two-level (school, student) Raw model was initially developed and set as a benchmark for the subsequent step, which involved adding the class level. This approach is consistent with the findings discussed in the previous chapter, which emphasized the importance of including all relevant levels to avoid bias in the models. Class can also contribute to variation in student academic outcomes (De Fraine et al., 2003; Mokonzi et al., 2020; Opdenakker & Van Damme, 2000; Tranmer & Steel, 2001). Therefore, a three-level (school, class, student) Raw model was developed and compared to the two-level Raw model to assess the significance of adding the class level.

In the subsequent stages, further analyses were performed to compare the Raw model with the addition of three more sophisticated models for four outcome measures. These models were developed to investigate school and class value-added effects using specific explanatory variables (see chapter 2 for model formulae). Table 5 presents the four models developed and the research questions addressed by the findings of each model.

Table 5

Multilevel Models developed in the study

Model	Fixed part explanatory variables	Description	Purpose
Raw model	Null	The null model that does not control for any explanatory variables	Estimate school differences in terms of 'raw' results in Total, Chinese, Maths, and English
VA model	Student prior attainment only (standardised JHSEE results in Total, Chinese, Maths, and	The prior attainment only model does not take student's particular background factors into	Estimate school differences after controlling for students' progress from the start of the junior high schooling to the end of junior high schooling

	English)	consideration	
CVA-1 model	Student prior attainment and significant variables of student background	CVA-1 model controls for student prior attainment and student background variables	Estimate school differences after controlling for student's progress over a period and background characterises. (Parents generally consider this Type-A effect when choosing a school (Bosker & Witziers, 1996; Raudenbush & Willms, 1996; Timmermans & Thomas, 2015)
CVA-2 model	Explanatory variables in CVA-1 model and significant contextual variables at school and class levels	CVA-2 model controls for CVA-1 variables and contextual data at school and class levels	Estimate school differences after controlling for student's progress over a period, individual student background variables, and contextual variables both at class and school level. (School officials generally consider this Type-B effect with the purpose of holding schools accountable for their students' results (Raudenbush & Willms, 1996; Timmermans & Thomas, 2015)

In short, this study examined school and class effects using four models, addressing sub-research questions 1.1-1.4. In statistical terms, a straightforward approach was taken to estimating school and class effects by specifying and calculating only the intercept residuals at school and class level. In other words, to answer RQ 1, this study has developed four random intercept models for the reasons discussed in chapter 2. A positive value-added score (intercept residual) indicated that a school may be performing above expectations compared to other schools in the sample, while a negative value-added score suggested that a school may be underperforming. To account for uncertainty in the estimation, a 95% confidence interval was employed (Thomas, 1998).

Sub-research question 1.5 sought to investigate the impact of using different value-added models on measuring school and class performance, given the differences in correlation between the school-level residuals from the Raw model to CVA-2 model. Moreover, school ranking positions were compared when using raw attainment measures versus VAMs, as well as the significant changes in school rank positions when different value-added models were employed. The overall findings were then used to address RQ1.

3.7.2 Thematic Analysis of Qualitative Data

3.7.2.1 The Advantages of Thematic Analysis

Thematic analysis was employed to interpret the interview data and to address RQ2 and sub-questions 2.1-2.5, which aimed to explore the perspectives of policymakers and headteachers on the potential benefits, disadvantages, and implementation of VAMs to support school evaluation and improvement. Thematic analysis is a systematic and flexible method of analysing qualitative data sets that enables researchers to generate rich and detailed accounts of the data while maintaining theoretical freedom, accommodating different research questions, sample sizes, data collection methods, and approaches to meaning generation (Braun & Clarke, 2006). Furthermore, it is a descriptive method that produces research findings that policymakers and practitioners find understandable (Howitt, 2019).

3.7.2.2 Thematic Analysis Procedures

This study adopted a hybrid approach to thematic analysis that combines a deductive a priori template of codes approach and the data-driven inductive approach (Fereday & Muir-Cochrane, 2006). In the deductive phase, an analytical framework was developed from the research questions, individual questions asked in the interviews, and previous literature in a similar field. For instance, the aims of school evaluation are associated with school accountability and improvement (OECD, 2009; 2020; Scheerens et al., 2003). Test-based school evaluation has advantages such as allowing comparisons and producing incentives, but it also has disadvantages such as reallocating efforts and teaching to the test (OECD, 2009; 2013). Value-added is linked to student progress and school contribution (e.g., OECD, 2008; Sammons et al., 2016; Scheerens et al., 2003). The implementation of value-added methods and potential influencing factors can be analyzed from various perspectives such as the political, methodological, and ethical (e.g., Ma, 2020; Munoz-Chereau et al., 2020; Saunders, 2001; Xin, 2020). This predetermined framework helped the researcher to develop codes and create more concrete themes from the interview data.

In the inductive approach, data-driven inductive coding analysis was used. Before conducting inductive coding analysis, interview data was transcribed into written text in Mandarin Chinese. The researcher then read and translated the written transcript from Mandarin Chinese into English. The

transcription stage is important because it familiarizes the researcher with the data and pushes the researcher towards understanding the data. Once that was done, the six critical phases of thematic analysis (Braun & Clarke, 2006) were adopted by the researcher as guidelines to analyze the interview data.

1) Data familiarization. This entails becoming familiar with the data by reading through the transcripts and searching for meanings and patterns.

2) Initial coding generation. At this stage, the codes describe lines of data rather than providing an insight or creativity.

3) Search for themes based on initial coding. All identical codes were aggregated together and categorized into several meaningful groups, which became the themes.

4) Review of themes. The themes were reviewed and evaluated to ensure accuracy. How the themes were interrelated with each other was examined and to what extent the overall themes aligned with the analytical framework.

5) Theme definition and labelling. Themes were defined and named, making sure they reflected the contents of the theme and could be distinguished from all the other themes.

6) Report writing. The final analysis was produced after revisiting and checking the themes. For example, checking whether the name of the themes fitted all the codes, and no coded sections would have been better suited to a different theme. Appendix 9 shows a table listing the final structure and labels of themes, codes and sub-codes and example key quote for each one.

3.8 Ethical Issues

The ethical considerations in this study were carefully reviewed and addressed. The guidelines from the School of Education (SoE), the British Educational Research Association (BERA) ethical guidelines for educational research (2018), and Chinese legal requirements were all taken into account. The University of Bristol's (UoB) research ethics procedure was followed, beginning with a discussion of the ethical issues with a colleague from UoB who was familiar with the study's context. The research was piloted after receiving ethical approval from UoB. The ethical issues that were considered in this

study, as outlined in the SoE ethics form, are discussed below.

3.8.1 Researcher Access

To recruit and enrol participants, the researcher obtained official permission from the LEA to conduct research in W district of G city in China. The participants included pupils aged 14/15, policymakers, and headteachers from public junior high schools in the district. The researcher worked closely with the LEA to obtain support for the study, including assistance in dispatching invitation letters to schools for the student questionnaires and for the interviews.

3.8.2 Informed Consent

During the quantitative phase, the researcher took great care to ensure the permission was obtained from the LEA. Therefore, the researcher sent a Request Permission Letter and obtained the permission from the LEA for the use of their administrative data and anonymised student examination data only identified via unique IDs for data matching (See Appendix 1). In regard to the student participants, informed consent was obtained for the collection of their background information. There were two main issues associated with the student online questionnaire, the first being the age of consent. According to Chinese legal requirements, consent can be obtained from a person aged 14 (GB/T 35273-2017 Information Technology – Personal Information Security Specification). After verifying with the LEA, it was confirmed that the students completing the questionnaires were 14/15 years old, making them eligible to give informed consent for their participation in this study. The second issue was the collection of separate signed consent forms from around 2,000 individual students. Under the General Data Protection Regulation (GDPR), creating active opt-in consent is deemed ethically acceptable. Therefore, the student's online questionnaire began with an information sheet, including a highlighted sentence that read, 'Completing the questionnaire indicates that I have read and agree with the research conditions outlined above,' at the bottom of the sheet. Once the students completed and returned the questionnaire, it indicated an active opt-in consent (see Appendix 2).

During the qualitative phase, informed consent was crucial to secure the trust of the interview participants. Participants were required to provide their consent via a signed consent form and agreement for the interview to be recorded before the interview started (see Appendix 3). The

researchers carefully considered any potential harms that could arise during the study and would continuously review the situation. Additionally, a complaint procedure was outlined in the study information sheet to ensure the participants' safety and protection.

3.8.3 Anonymity and Confidentiality

Ensuring anonymity and confidentiality is paramount in any research study. In this study, the researcher took various measures to guarantee anonymity and confidentiality of the research participants. In the case of quantitative data relevant to the students, including their JHSEE scores, SHSEE scores, and background information, the researcher removed any information that could potentially reveal the identity of individual students (Hennink et al., 2020). This was achieved by using the unique ID assigned to each student by the National Information Management System for K12 students, established in 2015 by the Ministry of Education in China. As all students registered for JHSEE, SHSEE, and school enrolment using their unique ID, the anonymity of students was ensured by using their unique ID in the questionnaire instead of identifiable information. The questionnaire data was then matched to the students' JHSEE and SHSEE scores.

In the case of interviewees, the researcher replaced their names with a capital letter during data analysis and research report phases. Additionally, the names of schools were replaced by an ID number, while the city and district were referred to as G city and W district, respectively. Both quantitative and qualitative data were stored in a password-protected account on the researcher's personal computer, following the UK Data Protection Act, to ensure their confidentiality.

3.8.4 Researcher as an Outsider

While the researcher possessed some familiarity with the study context and collaborated with the LEA to gain access to secondary data and collect primary data, it is important to note that the researcher is not a member of the LEA or the target schools, thus portraying an outsider perspective. This outsider position offers certain advantages, as it may allow for a fresh and impartial approach to the study context, free from biases or preconceived notions that insiders might carry. This objectivity might lend credibility to the research, ensuring that the findings are grounded in an unbiased assessment of the data. However, it is also essential to acknowledge the challenges associated with outsider positionality,

including difficulties in accessing data, building trust with interviewees, and accurately interpreting and contextualizing findings. To address these limitations, the researcher recognized the need for support from the LEA, given their proficiency in the local language and their understanding of the headteachers, which greatly facilitated effective communication in the interview and access to crucial data.

3.9 Research Quality

Although Creswell (2009) argues that utilizing the strengths of both qualitative and quantitative approaches is an advancement in mixed methods research, it is also acknowledged that there are potential threats to quality when using a mixed methods research design and analysis (Bryman, 2012). To address quality issues in this mixed-methods study, the validity and reliability standards of each approach were adopted as a perspective to assess quality, as recommended by several researchers (Creswell, 2009; Dellinger & Leech, 2007; Onwuegbuzie & Johnson, 2006; Tashakkori & Teddlie, 1998). The validity and reliability standards of both approaches are considered crucial for conducting mixed methods research across the different stages of research design, data collection, data analysis, and interpretation (Ihantola & Kihn, 2011).

In this study, the internal validity, external validity, and reliability of the quantitative work were assessed, while the credibility, transferability, and reliability of the qualitative work were considered (Lincoln & Guba, 1985). This approach ensures that the strengths of both quantitative and qualitative methods are leveraged while also addressing potential limitations and threats to quality in a rigorous manner.

3.9.1 Research Quality of Quantitative Part

3.9.1.1 Internal Validity

The concept of validity in quantitative research is related to the extent to which the study measures what it intends to measure. The primary inquiry in quantitative research regarding internal validity revolves around the ability to draw a sound conclusion from the utilized research design and implemented controls. Internal validity is affected when variables other than the independent variable influence the dependent variable, thus emphasizing the importance of statistical control variables in

research (Tashakkori & Teddlie, 1998). In this study, four types of multilevel value-added models were utilized to investigate the effects of schools and classes on students' academic achievements. The independent explanatory variables investigated in the fixed part of the multilevel models controlled for students' previous academic attainment, background characteristics, and contextual variables at the class and school levels. These contextual variables were derived from relevant VAMs research and were thoroughly discussed in the chapter on literature review. By controlling for these variables, the study aims to increase the internal validity of the results obtained from the multilevel models.

This approach of selecting dependent and independent variables and including them in multilevel models is consistent with the typical practice from Thomas and Mortimore (1996). By carefully selecting relevant and significant variables, the study avoids overloading the models with irrelevant factors, which can decrease the accuracy of the results. Therefore, all selected individually significant explanatory variables were then jointly inserted into multilevel models to determine which results were significant in terms of predicting the student's outcome achievement (Yu & Thomas, 2008). The variables that were not statistically significant were then excluded from the models. The exclusion of non-significant variables from the models helps to establish a consistent set of models that produce reliable and valid results. Finally, the consistent use of independent variables across outcome measures ensures that the comparison of school academic performance is valid and unbiased.

3.9.1.2 External Validity

External validity is the extent to which the findings of a study can be generalized to other samples, settings, and time periods. One of the major threats to external validity in quantitative studies is sampling bias or limitations. An insufficient sample size may not be representative of the entire population, and hence, generalizations to the target population cannot be made (Mohajan, 2017; Tashakkori & Teddlie, 1998).

In this study, data was collected from all public junior high schools in the district and the entire Grade 9 student cohort, which follows the recommendations of Willms (1992) that it is preferable to obtain data on whole cohorts of students for reducing sampling and measurement errors. The online questionnaire was distributed and administered by the teacher in charge of the class, resulting in a relatively high response rate of 80% (Rubin & Babbie, 2017) and a total of 1,596 valid responses out

of 1,745 returned questionnaires. Therefore, despite a relatively small sample compared to other value-added studies, by obtaining a relatively high response rate of 80%, the external validity of this study was enhanced.

However, other factors such as time and environmental issues may also affect external validity (Mohajan, 2017). As this study is limited to one LEA cohort over three years, its findings may not be generalizable to other time periods. Moreover, the estimates of school and class effects in W district may differ if the analysis were extended to a larger target population of Chinese public junior high schools. Furthermore, results obtained in one district may not be generalizable across different cities and provinces, as different districts are subject to varying education authorities. This may potentially impede the generalization of the research findings beyond the W district.

3.9.1.3 Reliability

In the quantitative approach, reliability refers to the consistency of a variable in measuring what it is intended to measure. A potential challenge to reliability during data collection is the lack of clear and standardized instructions. Vague item descriptions may lead to misinterpretation and, consequently, less reliable data (Ihantola & Kihn, 2011). To address this concern, Rubin and Babbie (2017) claim that researchers may utilize government-regulated measures, such as standardized test scores, or employ established measures that have demonstrated reliability in previous studies. Furthermore, researchers should conduct a pilot test of their research instruments to ensure their reliability and validity.

In this study, standard student examination scores were employed. Also, the development of the questionnaire was informed by previous international and local research, as discussed in the Section 3.5.1. The adequacy of the instructions given to students was assessed through piloting of the questionnaire. However, despite these efforts, some invalid questionnaires were still collected, where students either failed to answer some questions or provided inaccurate responses.

The decision to omit data was made based on specific criteria. For instance, some students did not answer questions about their parents' education and occupation, and potentially inaccurate responses may have been given in some cases regarding questions about items in their homes. To maintain data

quality and reliability, responses that fell below a certain threshold of completeness or consistency were omitted from the analysis. These issues pose a threat to the reliability of data collection to some extent, and the decision to omit certain data points aimed at ensuring the integrity and accuracy of the analysis.

3.9.2 Research Quality of Qualitative Parts

3.9.2.1 Credibility and Reliability

The qualitative approach is concerned with the credibility of the research, which is similar to the internal validity of the quantitative approach. Credibility refers to whether the research has accurately captured the phenomenon or attribute that is under investigation, and whether the findings reflect the participants' views and experiences in the research context (Bryman, 2012; Tashakkori & Teddlie, 1998). In contrast, reliability in the qualitative approach relates to the consistency of the findings and the extent to which they can be replicated (Tashakkori & Teddlie, 1998).

Threats to credibility in the qualitative approach may occur due to ambiguous or inaccurate interview questions, as well as problems with transcription and analysis (Ihantola & Kihn, 2011). To address these issues, participants in this study were informed of the purpose of the research and the anonymized use of their responses. They were also assured that there were no right or wrong answers, and that they did not need to provide a perfect response. Additionally, the researcher clarified any questions that participants did not understand during the interview, and participants were given the opportunity to review the interview data for potential biases. Finally, interview transcripts were given to each participant for their confirmation. By taking these steps, the credibility of the research was strengthened, and the risk of invalid findings was reduced.

3.9.2.2 Transferability

Transferability parallels external validity concerns about whether the research results can be extended to a broader context (Creswell & Poth, 2016). Threats to the transferability of a qualitative approach may occur if the researcher fails to reconnect the empirical research findings to other cases and theories and explain how an understanding of the research question was enhanced by new evidence (Golden-Bibble & Locke, 1993). Conducting this study may offer a contextual setting for the reform of school

evaluation methods in the context of China. The transferability of qualitative research findings is enhanced by providing a detailed description of the research context, sampling, data collection and analysis procedures, connecting the findings to existing theories and frameworks, and discussing the implications of the findings for the context under investigation and beyond.

3.9.3 Other Limitations

Conducting a mixed-methods study can be a complex process that requires significant time and resources. Therefore, it is challenging for a single Ed.D. researcher to collect and analyse both qualitative and quantitative data. In addition, accessing the participants and obtaining the students' standard examination results across different areas was challenging due to China's vast landscape and complex context. This leads to a limitation due to relatively limited sample size. Therefore, the generalisation and stability of the research findings were somewhat influenced.

Another limitation of this study is associated with the missing data. As outlined in section 3.5.1, although all 2006 students were expected to participate in the online questionnaire, not all completed it or provided valid information. 15% of students did not complete the online survey for various possible reasons, such as technical difficulties with the survey platform or internet connectivity issues, time constraints, or concerns about the privacy of their responses. An additional 5% of students completed the questionnaire but provided invalid answers, resulting in the omission of their data. While the average percentage of missing data is relatively low, the representativeness of the results may still be limited without conducting statistical tests to confirm the missingness mechanism.

Moreover, limitations exist in developing research instruments, especially the interview protocol. Although qualitative and mixed methods have been employed during the current development phase of SER, SER and VAMs research in China are still over-reliant on quantitative methods and data (Guo & Wang, 2021). As a result, the development of the interview schedule mainly referred to international literature while lacking adequate local Chinese literature for reference. Therefore, the scope and clarity of the interview themes may need to be improved.

Furthermore, the LEA plays a crucial role in supporting the collection of both quantitative and qualitative data. Their support is primarily driven by the value they place on the project findings for

future policy development. However, limitations may arise. For instance, even though interviews are conducted independently and confidentially, headteachers may still feel pressured to present their perceptions.

Lastly, it is important to consider the potential for researcher bias in this study. As noted by Brinkmann and Kvale (2005), researchers inevitably bring their personal experiences and roles to the research process, which can influence the dynamics of the study. In this study, the researcher possessed familiarity with the research context and had participated in prior research projects related to the research question. Thus, during the interviews, there may have been a risk of introducing bias, such as by nodding or sharing similar experiences with the participant, which could have influenced their responses in ways that aligned with the researcher's views. To mitigate this possibility, the researcher was diligent in reflecting on their potential biases throughout the study.

In short, it is not always feasible to eliminate all potential limitations or risks in research. However, what is crucial is to ensure that the limitations and risks that could affect the validity, reliability, and trustworthiness of the mixed methods employed in this study were clearly identified, minimized, and controlled to a reasonable extent. By acknowledging and addressing these limitations, the study's findings can be interpreted and generalised in a more informed manner. Furthermore, the study's transparency and comprehensive reporting of limitations and risks can serve as a valuable reference for future research in the field.

3.9 Conclusion

Overall, the research design employed in this study reflects a pragmatic approach. The sequential mixed method design is used to provide a deeper and more comprehensive understanding of the research problem and questions. The use of a student questionnaire in the first phase allowed for the collection of standardized data, which was used to build multilevel value-added models to address RQ1. In the second phase, semi-structured interviews were conducted with policymakers and headteachers to explore their perceptions of the potential implementation of value-added approaches in school evaluation practices, which addressed RQ2. Furthermore, ethical issues and issues related to validity, reliability, and transferability were discussed to ensure the rigor and trustworthiness of the study. Finally, limitations and potential biases were identified. Based on the research design, the

quantitative and qualitative findings will be illustrated in the following chapters.

Chapter 4 Quantitative Findings

4.1 Chapter Introduction

This chapter addresses RQ1, which investigates the range and extent of school and class performance across 11 public junior high schools located in the W district in Southwest China. To address this question, four multilevel models were utilized, as outlined in Chapter 3. The chapter begins with a descriptive analysis of the datasets employed to construct the models. Following that, the outcomes of four multilevel models are presented to estimate school and class effects. The subsequent section illustrates the changes in school performance when applying the four multilevel models. After addressing RQ1 and its five sub-questions, the chapter concludes by highlighting the primary findings derived from the quantitative phase.

4.2 Descriptive Analysis of Datasets

This study utilized outcome data on each student's Total, Chinese, Mathematics, and English examination results from the SHSEE, along with additional information about their individual characteristics. Specifically, the analysis incorporated information on the student's gender, prior attainment on entry to junior high school (assessed using the JHSEE examination in Total, Chinese, Mathematics, and English), whether the student received academic private tutoring, household items, items owned by the student, the number of books in the student's house, and the education and occupation of the student's parents. Additionally, the class and school compositional variables included the class and school mean total prior attainment in JHSEE. The following sections provide a detailed descriptive analysis of these variables.

4.2.1 Descriptive Analysis of JHSEE and SHSEE Datasets

This study collected data on student examination scores in 2018 JHSEE and 2021 SHSEE for each student in the form of raw marks. Both examinations were designed to meet the criterion of the National Curriculum Standards, and the scores for each examination were standardized (mean=0, standard deviation=1) to facilitate comparison of model findings, despite differences in full marks. Specifically, individual scores in Chinese, Mathematics, and English from both examinations were selected and matched over time for each student.

Table 6 summarizes the results of the JHSEE scores obtained by students in Grade 7 and the scores in Chinese, Mathematics, and English obtained in the SHSEE by the same cohort of students in Grade 9. The SHSEE serves as both an entrance examination for senior high school and a graduation examination for junior high school, making it compulsory for all students in Grade 9 to take the examination.

Table 6

Descriptive statistics of students' JHSEE and SHSEE raw scores

Variable	Valid Number	Mean	Std. Dev.	Min	Max	Std. Min.	Std. Max.
JHSEE (prior attainment)							
Chinese	1596	71.99	14.00	11.00	98.00	-4.36	1.86
Mathematics	1596	70.17	18.77	11.00	99.00	-3.15	1.54
English	1596	60.84	22.11	12.00	100.00	-2.21	1.77
Total score	1596	203.00	46.94	42.00	292.00	-3.43	1.90
SHSEE (outcomes)							
Chinese	1596	81.17	21.59	3.00	135.00	-3.62	2.49
Mathematics	1596	58.95	30.35	6.00	135.00	-1.74	2.51
English	1596	50.59	25.17	0.00	111.00	-2.01	2.40
Total score	1596	190.72	70.71	35.00	376.00	-2.20	2.62

Note. Total score is the sum of individual student scores in three subjects; Full Junior high school entrance exam score is 100 for each subject; Full Senior high school entrance exam scores are 150 for Chinese, Mathematics; 120 for English. Both examination scores were standardized (mean=0, standard deviation=1).

Table 6 illustrates that the English scores in JHSEE displayed the highest variability (22.11) compared to the scores in Mathematics (18.77) and Chinese (14.00). Conversely, in SHSEE, the Mathematics scores exhibited greater variability (30.35) than the scores in English (25.17) and Chinese (21.59). These findings align with previous research conducted in the province and other regions, which has found that the subject of English in JHSEE commonly exhibits the highest variability (Lv, 2015; Peng et al., 2013). However, the greater variability of Mathematics scores in SHSEE observed in this study is inconsistent with some other provinces or regions (Fan & Gao, 2019; Xin et al., 2012). Nevertheless, it is generally observed that Chinese scores in both JHSEE and SHSEE have the lowest variability among the three subjects.

4.2.2 Descriptive Analysis of Student Background Characteristics

A student questionnaire was developed to collect information about student background, as detailed in Appendix 2. However, only variables that were tested and selected for use in MLM analyses were included in Table 7, with the testing process demonstrated in Section 4.3.3.1. The variables included in the model were gender, whether the student had received academic curriculum-related tutoring outside of school in the past three years, and the number of books at home. Table 7 provides descriptive information about these variables.

Table 7

Selected student background variables

Student background variables	Number of valid values	Category	Valid percentage (%)
Gender	1596	Male	50.6
		Female	49.4
Academic private tutoring	1596	Yes	45.6
		No	54.4
Number of books at home	1596	0-10 books	4.6
		11-100 books	54.4
		101-200 books	29.2
		201 or above	11.8

4.3 RQ1: What are the range and extent of school and class effects of public junior high schools in the W district in Southwest China on student academic outcomes based on Raw, VA, CVA-1, and CVA-2 measures?

Tables 8-11 compare school and class effects on student outcome in SHSEE Total, Chinese, Mathematics, and English estimated by four value-added multilevel models, as outlined in Chapter 3. The fixed part variables represent controlled explanatory factors, while the random part examines variation in student outcomes at three levels: schools, classes within schools, and students within classes. The VPC estimates the impact of schools and classes on student outcomes by measuring the proportion of total variance attributed to these levels (Goldstein et al., 2002). Table 12 displays VPCs to schools, classes, and students for each developed model. The explanatory power or the "goodness of fit" of a model is assessed by determining how well explanatory factors account for variance in student outcomes (Muñoz-Chereau & Thomas, 2016; Strand, 1998). Hence, Table 12 also provides

information on the percentage of variance explained by the VA, CVA-1, and CVA-2 models, in comparison to the Raw model.

Table 8

Estimation of four multilevel models: SHSEE Total score

Fixe Part	Raw (RQ 1.1)		VA (RQ 1.2)		CVA-1(RQ 1.3)		CVA-2(RQ 1.4)	
Variables	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept	-0.14	0.15	-0.12	0.10	-0.33	0.12	-0.30	0.11
JHSEE Chinese	n/a	n/a	0.25**	0.02	0.22**	0.02	0.22**	0.02
JHSEE Mathematics	n/a	n/a	0.31**	0.02	0.31**	0.02	0.30**	0.02
JHSEE English	n/a	n/a	0.10**	0.02	0.08**	0.02	0.07**	0.02
Boy (vs. girl)	n/a	n/a	n/a	n/a	-0.15**	0.03	-0.15**	0.03
Receiving academic tutoring (vs. not receiving	n/a	n/a	n/a	n/a	0.17**	0.03	0.17**	0.03
Books (vs. 0-10 books)								
11-100 books	n/a	n/a	n/a	n/a	0.17*	0.07	0.17*	0.07
101-200 books	n/a	n/a	n/a	n/a	0.29**	0.08	0.28**	0.08
201 or above books	n/a	n/a	n/a	n/a	0.32**	0.08	0.31**	0.08
Class prior mean total attainment	n/a	n/a	n/a	n/a	n/a	n/a	0.25**	0.04
Random Part								
Between school: cons(intercept)	0.15	0.10	0.07	0.05	0.06	0.04	0.06	0.03
Between class: cons(intercept)	0.26	0.06	0.08	0.02	0.07	0.02	0.03	0.01
Between student: cons(intercept)	0.55	0.02	0.32	0.01	0.30	0.01	0.30	0.01
Units: School	11		11		11		11	
Units: Class	46		46		46		46	
Units: Students	1596		1596		1596		1596	
-2*loglikelihood:	3708.24		2824.74		2733.73		2701.99	

Note. Statistically significant at the *p<0.05; **p<0.001

Table 9

Estimation of four multilevel models: SHSEE Chinese

Fixe Part	**Row (RQ 1.1)**		**VA (RQ 1.2)**		**CVA-1(RQ 1.3)**		**CVA-2(RQ 1.4)**	
Variables	**Estimate**	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**
Intercept	-0.14	0.14	-0.11	0.09	-0.30	0.11	-0.28	0.11
JHSEE Chinese	n/a	n/a	0.41**	0.02	0.37**	0.02	0.37**	0.02
JHSEE Mathematics	n/a	n/a	0.22**	0.02	0.23**	0.02	0.22**	0.02
JHSEE English	n/a	n/a	0.03	0.02	0.01	0.02	0.01	0.02
Boy (vs. girl)	n/a	n/a	n/a	n/a	-0.25**	0.03	-0.24**	0.03
Receiving academic tutoring (vs. not receiving	n/a	n/a	n/a	n/a	0.09*	0.03	0.19*	0.03
Books (vs. 0-10 books)								
11-100 books	n/a	n/a	n/a	n/a	0.23*	0.08	0.23*	0.08
101-200 books	n/a	n/a	n/a	n/a	0.38**	0.08	0.37**	0.08
201 or above books	n/a	n/a	n/a	n/a	0.38**	0.09	0.38**	0.09
Class prior mean total attainment	n/a	n/a	n/a	n/a	n/a	n/a	0.15**	0.04
Random Part								
Between school: cons(intercept)	0.14	0.09	0.06	0.04	0.05	0.03	0.04	0.02
Between class: cons(intercept)	0.20	0.05	0.05	0.02	0.05	0.01	0.03	0.01
Between student: cons(intercept)	0.65	0.02	0.39	0.01	0.37	0.01	0.37	0.01
Units: School	11		11		11		11	
Units: Class	46		46		46		46	
Units: Students	1596		1596		1596		1596	
-2*loglikelihood:	3708.24		3135.52		3030.30		3017.56	

Note. Statistically significant at the *p<0.05; **p<0.001

Table 10

Estimation of four multilevel models: SHSEE Mathematics

Fixe Part	**Row (RQ 1.1)**		**VA (RQ 1.2)**		**CVA-1(RQ 1.3)**		**CVA-2(RQ 1.4)**	
Variables	**Estimate**	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**
Intercept	-0.12	0.14	-0.09	0.10	-0.32	0.12	-0.29	0.12
JHSEE Chinese	n/a	n/a	0.11**	0.02	0.10**	0.02	0.10**	0.02
JHSEE Mathematics	n/a	n/a	0.43**	0.02	0.41**	0.02	0.41**	0.02
JHSEE English	n/a	n/a	0.05*	0.02	0.04	0.02	0.03	0.02
Boy (vs. girl)	n/a	n/a	n/a	n/a	0.00	0.03	0.00	0.03
Receiving academic tutoring (vs. not receiving	n/a	n/a	n/a	n/a	0.15**	0.03	0.15**	0.03
Books (vs. 0-10 books)								
11-100 books	n/a	n/a	n/a	n/a	0.14	0.08	0.13	0.08
101-200 books	n/a	n/a	n/a	n/a	0.23*	0.08	0.22*	0.08
201 or above books	n/a	n/a	n/a	n/a	0.26**	0.09	0.25*	0.09
Class prior mean total attainment	n/a	n/a	n/a	n/a	n/a	n/a	0.25**	0.04
Random Part								
Between school: cons(intercept)	0.13	0.09	0.07	0.04	0.05	0.04	0.06	0.03
Between class: cons(intercept)	0.24	0.06	0.09	0.02	0.08	0.02	0.03	0.01
Between student: cons(intercept)	0.59	0.02	0.39	0.01	0.38	0.01	0.38	0.01
Units: School	11		11		11		11	
Units: Class	46		46		46		46	
Units: Students	1596		1596		1596		1596	
-2*loglikelihood:	3821.24		3128.58		3092.95		3066.89	

Note. Statistically significant at the *p<0.05; **p<0.001

Table 11

Estimation of four multilevel models: SHSEE English

Fixed Part	**Row (RQ 1.1)**		**VA (RQ 1.2)**		**CVA-1(RQ 1.3)**		**CVA-2(RQ 1.4)**	
Variables	**Estimate**	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**	**Estimate**	**S.E.**
Intercept	-0.14	0.14	-0.12	0.10	-0.27	0.12	-0.24	0.11
JHSEE Chinese	n/a	n/a	0.21**	0.02	0.18**	0.02	0.18**	0.02
JHSEE Mathematics	n/a	n/a	0.17**	0.02	0.17**	0.02	0.17**	0.02
JHSEE English	n/a	n/a	0.19**	0.02	0.17**	0.02	0.16**	0.02
Boy (vs. girl)	n/a	n/a	n/a	n/a	-0.12**	0.03	-0.12**	0.03
Receiving academic tutoring (vs. not receiving	n/a	n/a	n/a	n/a	0.22**	0.03	0.22**	0.03
Books (vs. 0-10 books)								
11-100 books	n/a	n/a	n/a	n/a	0.13	0.08	0.12	0.08
101-200 books	n/a	n/a	n/a	n/a	0.21*	0.09	0.20*	0.09
201 or above books	n/a	n/a	n/a	n/a	0.25*	0.09	0.24*	0.09
Class prior mean total attainment	n/a	n/a	n/a	n/a	n/a	n/a	0.27**	0.04
Random Part								
Between school: cons(intercept)	0.12	0.09	0.06	0.05	0.06	0.04	0.05	0.03
Between class: cons(intercept)	0.26	0.06	0.10	0.03	0.09	0.02	0.04	0.01
Between student: cons(intercept)	0.58	0.02	0.42	0.02	0.39	0.01	0.39	0.01
Units: School	11		11		11		11	
Units: Class	46		46		46		46	
Units: Students	1596		1596		1596		1596	
-2*loglikelihood:	3792.30		3240.17		3139.49		3110.92	

Note. Statistically significant at the *p<0.05; **p<0.001

Table 12

Percentage of variance attributable to the school, class, and student and percentage of variance explained

Outcomes and statistic	Level	Raw model %	VA model %	CVA-1 model %	CVA-2 model %
SHSEE Total score					
VPCs	Schools	15.37	15.32	13.69	14.36
	Classes	27.05	16.60	16.24	6.79
	Students	57.58	68.09	70.07	78.85
% variance explained	Schools	NA	50.68	59.59	62.33
	Classes	NA	69.65	72.76	89.88
	Students	NA	41.50	44.79	44.79
	Total	NA	50.53	54.63	59.68
SHSEE Chinese					
VPCs	Schools	13.98	11.88	10.75	8.22
	Classes	20.00	10.10	9.89	7.31
	Students	66.02	78.02	79.35	84.47
% variance explained	Schools	NA	56.20	63.50	73.72
	Classes	NA	73.98	76.53	83.67
	Students	NA	39.10	42.97	42.81
	Total	NA	48.47	52.55	55.31
SHSEE Mathematics					
VPCs	Schools	13.34	12.41	10.51	12.29
	Classes	24.68	15.74	15.56	7.20
	Students	61.97	71.85	73.93	80.51
% variance explained	Schools	NA	47.24	57.48	54.33
	Classes	NA	63.83	65.96	85.53
	Students	NA	34.24	35.59	35.59
	Total	NA	43.28	46.01	50.42
SHSEE English					
VPCs	Schools	12.73	11.03	10.28	10.25
	Classes	26.93	17.41	16.82	8.16
	Students	60.33	71.55	72.90	81.59
% variance explained	Schools	NA	47.54	54.92	59.84
	Classes	NA	60.85	65.12	84.88
	Students	NA	28.20	32.53	32.53
	Total	NA	39.46	44.15	50.10

Note. See text for an explanation of these measures

4.3.1 Sub-Question 1.1: *What is the estimated range and extent of school and class academic SHSEE performance (Total, Chinese, Mathematics, and English scores) in Chinese public junior high schools based on the Raw model?*

As described in Chapters 3, the first step in the MLM analysis is to establish the Raw model, which is subsequently expanded through three alternative models (VA, CVA-1, CVA-2) to test school and class effects using different explanatory data. In the Raw model, the dependent variables are standardized SHSEE outcome scores in Total, Chinese, Mathematics, and English, obtained by public junior high school students in 2021. The Raw model does not include any control for explanatory variables and serves to estimate the variability in schools' unadjusted performance. It serves as a base against which to compare the subsequent models and provides estimates of raw unadjusted school and class performance. Before analysing the findings, two Raw models were constructed, a two-level model (student, school) and a three-level model (student, class, school), and compared to demonstrate the superiority of the three-level model in subsequent statistical analyses (Kyriakides & Charalambous, 2004).

Table 13

Variance partition coefficients for the 2 and 3-level Raw models

Raw model (2-level)				
Fixed part	Total score (n=1596)	Chinese (n=1596)	Mathematics(n=1596)	English(n=1596)
Variable	Estimate (S.E)	Estimate (S.E)	Estimate (S.E)	Estimate (S.E)
Intercept	-0.194 (0.137)	-0.185 (0.139)	-0.153 (0.128)	-0.198 (0.128)
% of variance attribute to				
School level	19.73%	24.54%	17.11%	17.15%
Class level	NA	NA	NA	NA
Student level	80.27%	80.30%	82.89%	82.85%
-2*loglikelihood:	4161.067	4229.044	4213.473	4219.859
Raw model (3-level)				
Fixed part	Total score (n=1596)	Chinese (n=1596)	Mathematics(n=1596)	English(n=1596)
Variable	Estimate (S.E)	Estimate (S.E)	Estimate (S.E)	Estimate (S.E)
Intercept	-0.144 (0.151)	-0.142 (0.142)	-0.118 (0.143)	-0.137 (0.143)
% of variance attribute to				
School level	15.37%	13.98%	13.34%	12.37%
Class level	27.05%	20.00%	24.68%	26.93%
Student level	57.58%	66.02%	61.97%	60.33%
-2*loglikelihood:	3708.235	3958.427	3821.235	3792.299

A likelihood ratio test was conducted to compare the three-level model with the corresponding two-level model. The test measures the reduction in deviance (-2*log-likelihood) when transitioning from the two-level to the three-level model. The deviance reductions for SHSEE Total, Chinese, Mathematics, and English were 452.832, 270.617, 392.238, and 427.56, respectively. These values significantly exceeded the critical value of 3.84 (5% point of a chi-squared distribution on 1 degree of freedom is 3.84). Hence, the addition of the class level demonstrated a substantial improvement in the overall model fit (Leckie & Browne, 2021).

Table 13 displays that for the two-level Raw model, 19.73% of the variance in the total score is attributable to differences between schools, which can be interpreted as an estimate of the unadjusted (raw) school effect. The equivalent figures for Chinese, Mathematics, and English are 24.54%, 17.11%, and 17.15%, respectively. However, the three-level Raw model decomposes the total variance into separate school, class, and student components of variation. The results show that 15.37% of the variance in SHSEE total scores is attributable to differences between schools. The equivalent figures for Chinese, Mathematics, and English are 13.98%, 13.34%, and 12.37%, respectively. These findings suggest that the two-level Raw model may potentially overestimate the between-school variation. Therefore, the results obtained from the three-level multilevel models are particularly relevant and meaningful, despite the fact that two-level multilevel models were more frequently used in previous research on school effectiveness in China, as reviewed in Chapter 2. These findings are consistent with the results of Thomas (2020) and Peng et al. (2013), while being significantly smaller than the results of Du and Yang (2011), Fan and Gao (2019), and Xin et al. (2012).

In short, these results suggest that there are significant differences between schools in terms of their impact on student outcomes. However, it is important to note that most of the variance in student outcomes is still attributable to differences within schools and between classes and students, rather than between schools. This highlights the importance of considering the clustering of students within classes, the composition of classes within schools and the individual characteristics of students when examining school effectiveness. Nevertheless, Raw models are the starting point in the analysis of schools' performance and provide new estimates of unadjusted (raw) school and class effects in the local context of this study. However, they are not value-added models, since students' progress is not

being measured and intake differences are not adjusted for (Goldstein, 1997).

4.3.2 Sub-Question 1.2: *What is the estimated range and extent of school and class academic SHSEE performance (Total, Chinese, Mathematics, and English scores) in Chinese public junior high schools based on the VA model*?

Given that schools may have students with varying levels of prior attainment, it is essential to investigate whether school outcomes differ even after accounting for these differences in student intake. Controlling for student prior attainment is widely recognized as a necessary step when assessing school performance (Marks, 2017). Therefore, to address Sub-question 1.2, the VA models that incorporated only the three measures of prior attainment: JHSEE scores for Chinese, Mathematics, and English were utilized (see Tables 8-11).

4.3.2.1 School and Class Effects

After controlling for student prior attainment in the fixed part of the models, the results indicate a reduction in the variance across the four student outcomes that can be attributed to differences between schools, from the Raw models (which ranged from 12.73% to 15.37%) to the VA models (which ranged from 11.03% to 15.32%). A more significant reduction was observed in the variance across the four student outcomes that can be attributed to differences between classes within schools, from the Raw models (which ranged from 20.00% to 27.05%) to the VA models (which ranged from 10.10% to 17.41%), compared to the reduction in differences between schools. This suggests not only that student prior attainment should be considered when evaluating school and class performance but also that classes within schools differ in their intake factors.

The results also indicate that subject-specific differences play a significant role in evaluating schools' academic performance. Specifically, the analysis shows that school differences in the effects on students' performance in SHSEE Mathematics are more pronounced than those in the other two subjects. On the other hand, variations between classes in their effects on students' performance in SHSEE English are greater than those observed in the other two subjects. These findings suggest that evaluating schools' performance based solely on total SHSEE scores may not be sufficient, and class-level data should be considered to gain a more comprehensive understanding of academic performance.

The importance of including subject-specific data and class-level information in the evaluation of schools' academic performance has also been emphasized in previous studies (Thomas et al., 1997).

4.3.2.2 Testing for School Effects and the "Goodness of Fit" of the VA Models

The results of the likelihood ratio testing reveal a substantial reduction in deviance (−2*log-likelihood) of 883.5, 822.91, 692.65, and 552.13 points for SHSEE total, Chinese, Mathematics, and English, respectively, when moving from the Raw model to the VA model. These values greatly exceed the critical value of 3.84, indicating a statistically significant reduction in school effects. Therefore, the VA model is preferred over the Raw model (Leckie & Browne, 2021). Additionally, the inclusion of student prior attainment variables explains 50.68% of the school variance and 50.53% of the total variance in SHSEE total scores. For SHSEE Chinese, Mathematics, and English scores, the corresponding values are 56.20% and 48.47%, 47.24% and 43.28%, and 47.54% and 39.46%, respectively. These findings suggest that adjusting for prior attainment is crucial, as emphasized by Leckie and Prior (2022), Thomas (2001), and Zhang (2016).

In summary, the results of the VA model demonstrate that students' prior attainment has a significant and positive impact on their SHSEE scores in Chinese, Mathematics, and English. However, it appears that this factor alone is insufficient in explaining the variance in student outcomes. Previous research suggests that demographic and socioeconomic backgrounds are also important predictors of student achievement and should be considered when examining school performance. Therefore, to address this issue, the CVA-1 model is developed, which includes both prior attainment and student background factors.

4.3.3 Sub-Question 1.3: *What is the estimated range and extent of school and class academic SHSEE performance (Total, Chinese, Mathematics, and English scores) in Chinese public junior high schools based on the CVA-1 model*?

4.3.3.1 Developing CVA-1 Model

The initial step in developing the CVA-1 model was to identify significant student background variables involved in the CVA-1 model. According to previous studies by Thomas (2001), Thomas and Mortimore (1996), and Munoz-Chereau (2013), this study examined each relevant student background

variable obtained from the questionnaire individually with the Raw model, to assess the statistical significance (at the $p<0.05$ level) of each variable in explaining one or more SHSEE outcomes. Subsequently, all the student background variables that were individually significant were jointly analysed in the Raw model to determine which variables were overall significant for one or more SHSEE outcomes. Eventually, the jointly significant variables were included in the CVA-1 model, while the statistically non-significant variables were excluded. By eliminating non-significant variables, only three student background variables (gender, academic curriculum tutoring outside of school, and number of books in the household) were found to be significant (at the $p<0.05$ level) and were included in the CVA-1 model.

It should be noted that although the mother's and father's education qualifications and occupations were statistically significant when examined individually, they were not significant when analyzed jointly in the CVA-1 model. One possible explanation for this could be the confounding factor of students receiving private academic tuition. Guan (2022) notes that parents' educational qualification level and family income level significantly increase the likelihood of students receiving private academic tutoring. Thus, these variables might be confounded. Another explanation for this finding could be related to the inclusion of prior attainment in the joint examination. As discussed in Chapter 2, student background characteristics have a diminished role compared to student prior attainment due to confounding. For example, Ballou et al. (2004) reported that some student background variables may become less significant when a rich set of prior attainment measures is used.

4.3.3.2 School and Class Effects

After adjusting for above significant student background variables in the fixed part of the CVA-1 models, there was a further reduction in the variance across the four student outcomes that can be attributed to differences between schools, from the VA models (ranging from 11.03% to 15.32%) to the CVA-1 models (ranging from 10.28% to 13.69%). A similar range of reduction was found in the variance across the four student outcomes that can be attributed to differences between classes within schools, from the VA models (ranging from 10.10% to 17.41%) to the CVA-1 models (ranging from 9.89% to 16.82%). These findings suggest that the CVA-1 model has to some extent successfully accounted for a significant proportion of the variation in student outcomes that was previously

attributed to school and class effects. However, the remaining variation may be explained by other factors that were not included in the model, such as unmeasured student background characteristics, teaching quality, and other school-level factors. It is also worth noting that although the model underlines the importance of considering student background characteristics and prior attainment when evaluating school and class effectiveness, the resulting school and class effects change relatively little when compared with the change from the Raw model to the VA model.

With regards to subject differences, the analysis revealed that the variations between schools in their effects on students' performance were similar across the three subjects. Specifically, the percentage of variance across schools was found to be 10.75%, 10.51%, and 10.28% for SHSEE Chinese, Mathematics, and English, respectively. Consistent with the VA model, the variation between classes in their effects on students' performance in English was greater than that observed in the other two subjects. This finding may suggest that English SHSEE scores are more influenced by class-level factors compared to Chinese and Mathematics SHSEE scores. Notably, the results also showed that Chinese SHSEE scores are more influenced by school-level factors, as evidenced by a larger proportion of the variance in student Chinese SHSEE scores attributable to the school level than that at the class level. This finding suggests that improving student performance in Chinese should focus more on addressing school-level factors, indicating the importance of school-wide interventions in promoting academic achievement in this subject.

4.3.3.3 Testing for School Effects and the "Goodness of Fit" of CVA-1 Models

After conducting the likelihood ratio testing to compare the VA model and the CVA-1 model, the results showed a significant reduction in deviance for all four student outcomes - SHSEE total, Chinese, Mathematics, and English - in the CVA-1 model as compared to the VA model. The reduction in deviance exceeded the critical value of 3.84, indicating statistical significance of school effects in the CVA-1 model. After accounting for student background factors in addition to prior attainment variables, the CVA-1 model explained 54.92-63.50% of the school variance and 44.15-54.63% of the total variance in students' SHSEE scores. These results indicate that the inclusion of student background factors further improved the model's goodness of fit and supported the importance of accounting for both prior attainment and student background factors when evaluating school effectiveness. These

findings are consistent with previous studies by Leckie and Goldstein (2019) and Thomas (2001).

It is crucial to acknowledge that despite the reductions in variance at the school and class levels, school and class effects remain present even after additionally accounting for student background factors. As demonstrated in Table 12, a portion of the unexplained variance across four outcomes, ranging from 10.28% to 13.69%, was attributed to differences between schools, while another portion, ranging from 9.89% to 16.82%, was attributed to differences between classes within schools. Therefore, to explore a more accurate evaluation of school and class performance, a CVA-2 model was developed in the subsequent section. This model involved additional adjustments for class mean prior total attainment, which may have the potential to provide a more refined analysis of school and class effectiveness.

4.3.4 Sub-Question 1.4: *What is the estimated range and extent of school and class academic SHSEE performance (Total, Chinese, Mathematics, and English scores) in Chinese public junior high schools based on the CVA-2 model?*

4.3.4.1 Developing CVA-2 Model

To explore more precise estimates of school and class performances, the CVA-2 model was developed by additionally adjusting for school and class mean prior total attainment. When school mean prior total attainment alone is additionally adjusted for in the CVA-2 model, it shows a positive statistically significant relationship. However, this relationship became non-significant when class mean prior total attainment was also adjusted. These findings are consistent with previous research by Leckie and Prior (2022), Marks (2021), and Thomas and Mortimore (1996), who have also found that once appropriate adjustments are made at the student and class levels, the overall context of the school may have little impact on the variation in student outcomes. Moreover, Moknzi et al. (2020) have shown that, in addition to individual variables, the class variable in terms of average pre-test scores is a good predictor of student achievement and explains a considerable proportion of the residual variance between schools and classes within schools. Thus, the CVA-2 model in this study was developed with the additional adjustment for class mean prior total attainment to estimate school and class performances.

4.3.4.2 School and Class Effects

Although the fixed part of the CVA-2 model was additionally adjusted for class mean prior total

attainment, the adjustment made a relatively small reduction to the variance across student outcomes in SHSEE total, Mathematics, and English scores that is attributable to differences between schools. Only for SHSEE Chinese, there was a slightly further reduction in the percentage of the variance in SHSEE Chinese scores that is attributable to differences between schools (from 10.75% to 8.22%). However, the findings indicate that controlling for class mean prior total attainment had a significant impact on reducing the variance in SHSEE total, Mathematics, and English scores that is attributable to differences between classes within schools (from 16.24% to 6.79%, 15.56% to 7.20%, 16.82% to 8.16% for SHSEE total, Mathematics, and English scores respectively). This suggests that classes within schools may differ by intake factors, especially the class mean prior total attainment found in this study. In other words, there may be non-random allocation of students to classes based on their prior attainment. Students who with higher or lower prior attainment levels may be grouped together in a class, which may be intentional or unintentional, however, it is likely to have a significant impact on student outcomes. It is worth noting that class mean prior total attainment may also be associated with the quality of teaching. For example, the best teachers may be allocated to the class with the highest mean prior total attainment (Timmermans & Thomas, 2015). Therefore, controlling for class mean prior total attainment may be appropriate for some evaluation purposes, such as school internal accountability or teaching effectiveness evaluation.

Overall, the findings suggest that peer effects at the school level may have little practical importance in this study, while the effects at the class level require greater attention. However, it is important to note that the lack of significance of school mean prior total attainment in this study may be attributed to the small number of schools included. Therefore, future research with a larger sample size is necessary to test the significance of school prior mean attainment, as some previous studies have demonstrated positive effects of school-level prior attainment (Marks, 2021; Muñoz-Chereau, 2013; Muñoz-Chereau & Thomas, 2016).

4.3.4.3 Testing for School Effects and the "Goodness of Fit" of CVA-2 Models

The results of the likelihood ratio testing indicate that the CVA-2 model is preferred to the CVA-1 model, as the former model shows a significant reduction in deviance in all outcomes (−2*log-likelihood) of 31.73, 12.74, 26.06, and 28.57 points in SHSEE total, Chinese, Mathematics, and

English respectively. Thus, the school effects are statistically significant, and the CVA-2 model is preferred to the CVA-1 model (Leckie & Browne, 2021). Moreover, the additional inclusion of class mean prior total attainment in the CVA-2 model improves the goodness of fit, as it explains a larger proportion of the variance in student outcomes (50.10-50.68% of the total variance and 54.33-73.72% of the school variance in students' SHSEE scores was explained) (see Table 11). The "goodness of fit" of the CVA-2 model, in this sense, was further improved when compared with the CVA-1 model. These findings are consistent with previous studies by Munoz-Chereau (2013), Munoz-Chereau and Thomas (2016), Peng et al. (2013), Salim (2011), and Thomas and Peng (2011).

In summary, after controlling for student prior attainment, student background factors, and class mean prior total attainment, of the remaining total variance, 8.22-14.36% was attributable to differences between schools, and 6.79-8.16% was attributable to differences between classes, thereby demonstrating the school and class effects in the local context of this study. In comparison with previous research on similar subjects, such as studies conducted by Munoz-Chereau & Thomas (2016), Thomas (2001; 2020), and Thomas & Mortimore (1996), the current study found similar and slightly higher results in terms of the range of school effects. It is important to note, however, that a random slope model may provide a more powerful estimation of school and class effects than the random intercept model used in this study. Future studies could potentially benefit from employing a random slope model to investigate school and class effects further.

4.3.5 Sub-Question 1.5: *Do Differential School and Class Effects on Three Subjects Exist?*

Based on the above discussion, the CVA-2 model appears to be the optimal model for estimating school and class effects, as it has demonstrated both statistical significance and better model fit compared to the CVA-1 model. To further investigate the school and class effects on promoting students' SHSEE scores, the CVA-2 models were used to estimate the differential effects on the three individual subjects. The results of these analyses are presented in Tables 14 and 15. It is worth noting that only the intercepts were allowed to vary between schools to provide a single effectiveness measure, as previously stated. These tables provide comprehensive information on the generalizability of the school and class effects on promoting student achievement.

Table 14

Correlations between schools' effects on SHSEE Chinese, Mathematics, English for 11 schools(CVA-2 Model)

		Chinese	Mathematics	English
Chinese	Pearson Correlation	1	.799**	.732*
	Sig. (2-tailed)		0.003	0.010
Mathematics	Pearson Correlation	.799**	1	.745**
	Sig. (2-tailed)	0.003		0.009
English	Pearson Correlation	.732*	.745**	1
	Sig. (2-tailed)	0.010	0.009	

Table 15

Correlations between classes' effects on SHSEE Chinese, Mathematics, English for 46 classes(CVA-2 Model)

		Chinese	Mathematics	English
Chinese	Pearson Correlation	1	.541**	.476**
	Sig. (2-tailed)		0.000	0.001
Mathematics	Pearson Correlation	.541**	1	.408**
	Sig. (2-tailed)	0.000		0.005
English	Pearson Correlation	.476**	.408**	1
	Sig. (2-tailed)	0.001	0.005	

Note. **. Correlations is significant at the 0.01 level (2-tailed); *. Correlation is significant at the 0.05 level (2-tailed)

Table 14 presents strong positive correlations between the effects of schools on student SHSEE performance in Chinese, Mathematics, and English. The correlations observed are relatively higher than those reported in previous studies (e.g., Du & Yang, 2011; Peng et al., 2006; Thomas et al., 1997). Notably, the correlation coefficient of 0.799 between school effects on Chinese and Mathematics indicates a stronger positive relationship between these two subjects than between the other pairs of subjects. This implies that schools that perform well in Chinese subject are more likely to perform well in Mathematics subject as well. These similarities in departmental effects within schools may be attributed to various factors such as school policies, department planning, and educational resources allocated to the subjects. However, it should be noted that while the findings suggest an overall relatively strong correlation between school effects on three subjects, they do not necessarily prove that schools that perform well in one subject area will perform well in the other two areas as well. There may be other factors that impact students' performance in each subject, such as motivation, private tutoring, and individual differences in learning abilities.

Table 15 illustrates the moderate positive correlations between class effects on SHSEE performance

in Chinese, Mathematics, and English. Notably, the class effects on Chinese are most closely correlated with those on Mathematics, whereas the correlations for English and the other two subjects are relatively low. The results for class effects on the three subjects indicate that there is no significant similarity in departmental results, and differences exist within some classes. The discrepancies may be due to teacher quality, teaching practices, and the differential allocation of resources in the classes. For instance, certain classes may have more experienced teachers, which can enable students to perform better in specific subjects. The weak correlation between class effects on English and each of the other two subjects suggests that the teaching strategies and resources for English may not be commonly shared with the other two subjects.

It's worth noting that the CVA-2 model revealed gender disparities in three subjects. Specifically, statistically significant negative effects of student gender (with boys compared to girls) were observed in Chinese and English SHSEE outcomes. However, in the case of Mathematics, the impact of student gender on SHSEE outcomes was found to be non-significant. This implies the importance of addressing gender disparities in the subjects like Chinese and English, where boys appear to face greater challenges compared to girls in achieving favourable SHSEE outcomes, consistent with previous research indicating that girls outperform boys in various attainment measures (OECD, 2008). Conversely, the lack of a significant gender gap in Mathematics differs from some performance comparison findings suggesting boys' stronger performance in this subject (OECD, 2008). This underscores the importance of further investigation into gender disparities in specific subjects, which can inform targeted education policies and decisions aimed at addressing these disparities.

In summary, the findings suggest that a narrow focus on total examination scores or individual subject scores may obscure important information regarding a school's overall performance. Additionally, schools may have similar resources and policies across various subjects, but class-level analyses reveal that differences in teaching practices, resources, and student learning exist. These subject-specific differences underscore the importance of considering individual subject performance in addition to total scores when evaluating a school's effectiveness.

4.3.6 Sub-Question 1.6: *How do school rankings change across models?*

4.3.6.1 Correlations Between the School-level Residuals from the Raw Model to CVA-2 Model

To compare the four multilevel models used in this study, a detailed analysis of their practical differences is presented in this section. One way to compare these models is to examine the correlations of the school-level residuals between them. The degree of correlation between the two models may affect the resulting school rankings. If the two measures are highly correlated, the school rankings generated would be similar. Conversely, if the correlation is low, the school rankings may differ significantly (Leckie & Goldstein, 2017).

Table 16

Pearson Correlation (Sig. (2-tailed)) *between the school-level residuals from raw model to CVA-2 model*

SHSEE Total score	Raw	VA	CVA-1	CVA-2
Raw	1	.930**	.918**	0.467
VA	.930**	1	.999**	.756**
CVA-1	.918**	.999**	1	.777**
CVA-2	.467	.756**	.777**	1
SHSEE Chinese	Raw	VA	CVA-1	CVA-2
Raw	1	.923**	.910**	.702*
VA	.923**	1	.998**	.920**
CVA-1	.910**	.998**	1	.932**
CVA-2	.702*	.920**	.932**	1
SHSEE Mathematics	Raw	VA	CVA-1	CVA-2
Raw	1	.929**	.917**	0.474
VA	.929**	1	.999**	.756**
CVA-1	.917**	.999**	1	.780**
CVA-2	.474	.756**	.780**	1
SHSEE English	Raw	VA	CVA-1	CVA-2
Raw	1	.945**	.934**	0.480
VA	.945**	1	.998**	.739**
CVA-1	.934**	.998**	1	.760**
CVA-2	.480	.739**	.760**	1

Note. **. Correlations is significant at the 0.01 level (2-tailed); *. Correlation is significant at the 0.05 level (2-tailed)

Table 16 presents the correlations between various models and four outcomes. Notably, the lowest or insignificant correlation was observed between the Raw model and the CVA-2 model, as these models differ significantly in their adjustments. This finding is consistent with prior research, such as Leckie

and Prior (2022) and Leckie and Goldstein (2019), as well as Thomas and Mortimore (1996). Furthermore, a high correlation of 0.92 was observed between the Raw model and the VA model, indicating that schools with the highest raw mean achievement may still appear the most effective after adjusting for student prior attainment. This finding is aligned with the conclusions of Leckie and Prior (2022) and Munoz-Chereau (2013).

Regarding the relation between VA and CVA-1, a considerable closeness was observed, indicating that adjusting for student sociodemographic characteristics has relatively little impact on the relative performance of most schools. This could be attributed to the reduced role of student background characteristics, as discussed in Section 4.3.3.1. In terms of comparing CVA-1 and CVA-2 school effects, a low correlation of 0.77 was observed across SHSEE total, Mathematics, and English scores, while a high correlation of 0.93 was found for Chinese. These results suggest that adjusting for class mean prior total attainment has a greater impact on the relative performance of most schools, particularly for SHSEE total, Mathematics, and English scores. However, this impact is relatively small for Chinese scores.

In short, the findings suggest that both student prior attainment and class mean prior total attainment seem to be the key variables that impact the estimation of school residuals in this study. The use of different multilevel models led to changes in school effects. However, it is crucial to note that the sample schools in this study may have relatively homogeneous characteristics to some extent. Therefore, the relationships established in this study cannot be assumed to hold true in different contexts, and further research is needed to validate the findings (OECD, 2008).

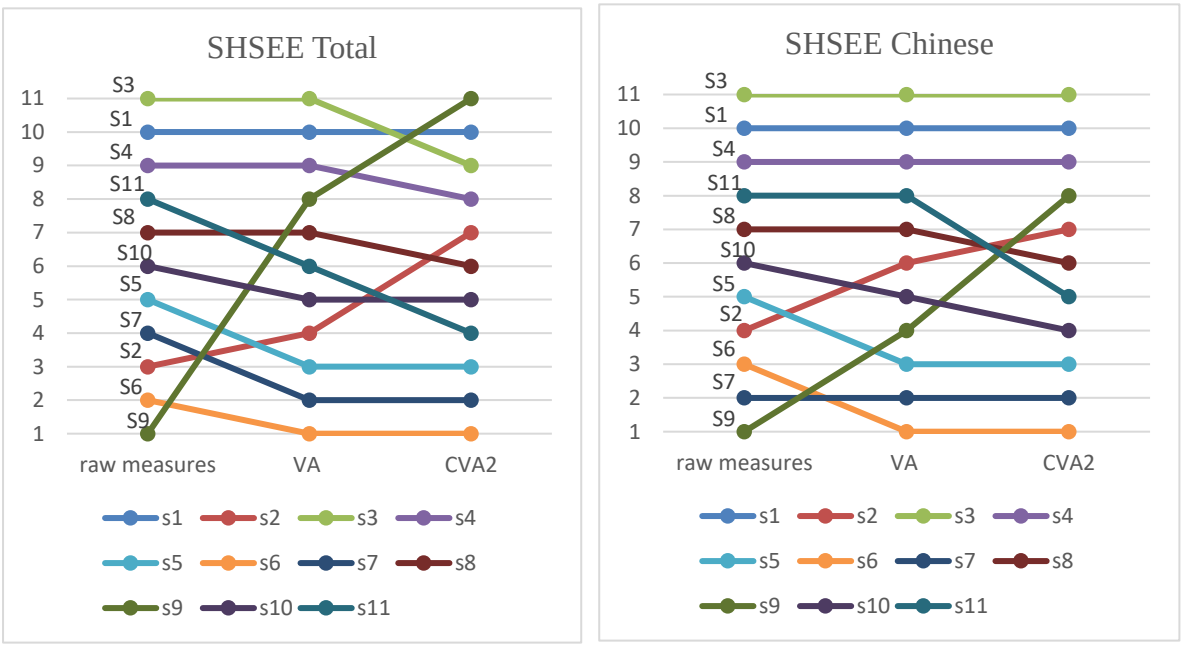
4.3.6.2 Changes in School Performance Rankings Based on Raw and Value-added Scores

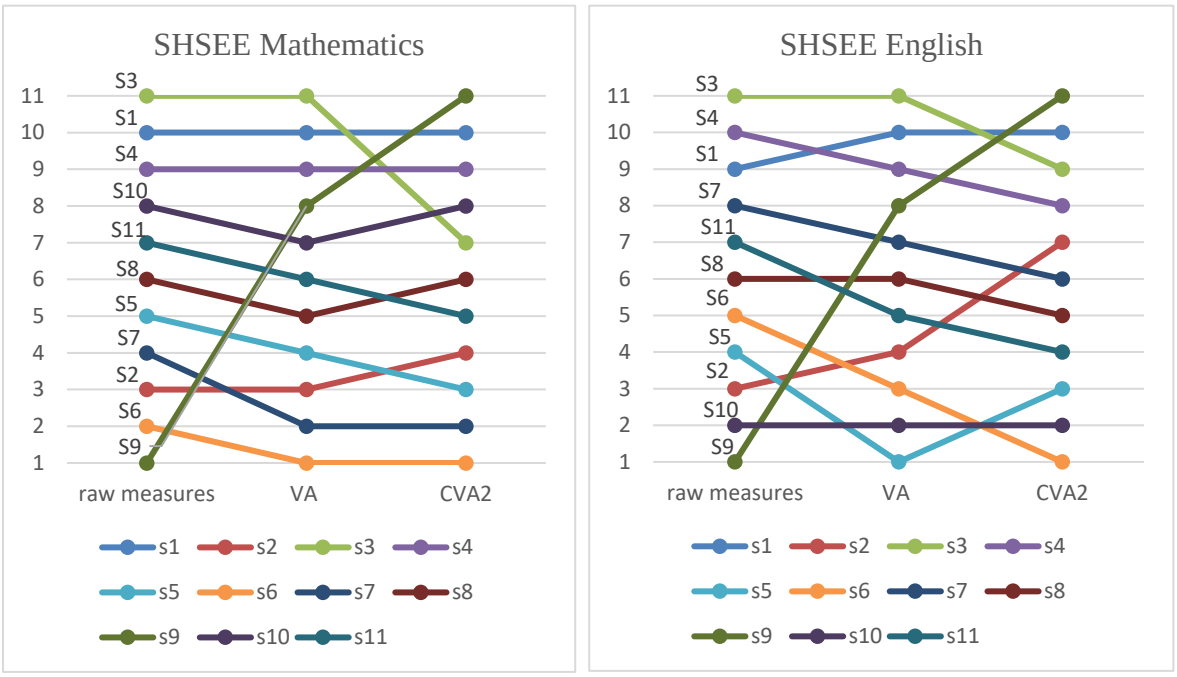
This section presents the differences between raw attainment measures and VAMs in ranking schools. However, it is crucial to interpret value-added scores cautiously. In particular, the confidence intervals are essential to assess whether a school's performance is statistically significant. If the confidence interval for a school's residual does not overlap with zero, this suggests that the school's value-added score is statistically significantly different at the chosen level of confidence (Leckie& Goldstein, 2019; Thomas, 1998). However, regarding the findings of this study, some schools' value-added scores may not be statistically significant. Despite that, presenting the changes in school performance rankings in

this study has several reasons. First, the presentation of school ranks based on both raw attainment measures and VAMs can highlight the differences between the two measures. Moreover, presenting the rank changes in figures may help to make the comparison more accessible and easier to interpret for audiences, such as policymakers or school administrators. Additionally, the findings in this study are mainly exploratory and could be seen as a pilot for larger VAMs studies in China. Therefore, based on the findings in section 4.3.6.1, the comparison is based on the ranks of school raw mean scores, value-added scores from the VA and CVA-2 models across four outcomes.

Figure 3

Changes in school rank positions in SHSEE total, Chinese, Mathematics, English scores





Note. S1-11 present the school ID. The horizontal lines denote the school performance base on raw SHSEE school mean scores and value-added scores on VA and CVA-2 models. The vertical lines denote the rank positions of schools in a descending order (rank 1 indicates the lowest level of performance) .

As shown in the figure, approximately half of the schools experienced rank changes when the VA model was used instead of raw SHSEE school mean total scores. Notably, some schools exhibited significant performance differences when student prior attainment was adjusted in the VA model. For instance, School 9's rank positions increased significantly in SHSEE total, Mathematics, and English. In contrast, the rank positions of Schools such as 5, 7, and 11 experienced a slight decline. However, the ranking of Schools 1, 3, and 4 remained relatively stable when student prior attainment was considered in the VA model. This is in line with the above analysis that schools with the higher mean current achievement still tend to the schools appeared more effective once student prior attainment is adjusted for.

The comparison of school rankings between the VA and CVA-2 models highlights the substantial impact of the model choice on school performance. More than half of the schools experienced changes in their rank positions from the VA model to the CVA-2 model, indicating that student background characteristics and class mean prior total attainment may have a considerable impact on school performance. School 9 was particularly noteworthy in its remarkable progress in four outcomes. When

controlling for student background factors and class prior mean total attainment, School 2's performance in SHSEE English was outstanding, with a rank position increase of four places. Conversely, the rank position of School 11 in SHSEE Chinese decreased dramatically in the CVA-2 model, while it did not change when student prior attainment was adjusted in the VA model.

In short, the presented figures demonstrate that the evaluation methods employed can yield substantial variations in school performance rankings. While the use of school residuals for ranking is subject to limitations, the graphical representation of individual school performance changes has the potential to facilitate the identification of schools that have made progress and those that require additional support or attention. This approach may prove particularly useful for non-statistically skilled audiences seeking to understand the difference between raw attainment measures and value-added measures and the effect of different value-added models. By illustrating the performance changes of individual schools, policymakers and school administrators may gain insights into the effects of educational interventions and identify areas for targeted improvements. However, further large size studies are needed to validate the findings of this study in different contexts.

4.4 Chapter Conclusion

In conclusion, the findings of this chapter highlight the importance of using appropriate evaluation methods to measure school and class effectiveness accurately. The results demonstrate the limitations of relying solely on raw SHSEE results to assess school performance and the potential problems that may arise from incorrect interpretations of school effectiveness. However, it is important to note that policymakers may make different choices based on multiple reasons (Leckie & Prior, 2022), and school staff may have different perceptions of the information provided by the value-added models (Saunders, 2001). Therefore, in the next chapter the perceptions of policymakers and school headteachers about the potential benefits, disadvantages, and implementation of VAMs to support school evaluation and improvement will be explored.

Chapter 5 Qualitative Findings

5.1 Chapter Introduction

This chapter addresses RQ2, which explores stakeholder perspectives regarding the potential benefit, disadvantages, and implementation of VAMs in school evaluation practices when local contexts are considered. To address this question, interviews were conducted with a selected sample of three policymakers from the LEA and three headteachers from public junior high schools in the sample district, as outlined in Chapter 3. The chapter begins with a brief introduction to the interview participants involved in the study. Subsequently, the analysis presents five themes that were derived from the interview data. The first theme, titled " Purpose of School Evaluation," maps the interviewees' perspectives on why schools engage in evaluation activities and what they hope to achieve through these activities. The second theme, " Advantages and Disadvantages of the Unadjusted Raw attainment-based School Performance Measures," captures the current local school evaluation practices that are based on the unadjusted raw attainment measures. The third theme, “The Concept of Value-Added”, maps the level of understanding demonstrated by the interviewees with respect to value-added and VAMs. The fourth theme, “Motivations to Implement VAMs in Practice”, reveals the driving factors that influence the interviewees’ decisions to adopt VAMs in school evaluation practices. Lastly, the fifth theme, “Factors that Influence the Implementation of VAMs”, sheds light on the contextual factors that potentially impact the implementation of VAMs (for a detailed examination of the themes and codes, refer to Appendix 9). Upon thorough exploration of RQ2 and its sub-questions, the chapter concludes by emphasizing the primary findings generated from the qualitative phase of the study.

5.2 Interviewees

In accordance with the information presented in Section 3.5.2, a total of six local stakeholders participated in voluntary interviews, comprising three policymakers and three school headteachers. To clarity and brevity in the subsequent analysis, acronyms were employed to refer to these interviewees. Table 4 (page 55) contains detailed information about these six interviewees.

5.3 RQ2: What are stakeholders' perceptions of the potential advantages, disadvantages, and implementation of the value-added approach in school evaluation practices when local contexts are considered?

5.3.1 Sub-Question 2.1 *What are interviewee perceptions on the purpose of current junior high school academic performance evaluation?*

Previous studies have investigated the purposes of school evaluation. For example, OECD (2008) argued that accurate measurement of school performance can achieve policy objectives such as evaluating investment in schools, assessing the quality of education provided by a school, and identifying areas for improvement. Similarly, Scheerens et al. (2003) emphasized that the motivations behind school evaluation include formally regulating desired levels of educational quality outcomes, holding schools accountable, and supporting ongoing improvement in education. In some education systems, school performance evaluation also aims to provide information for school choice, self-evaluation, and target-setting (OECD, 2008). In comparison to previous findings, the responses of participants in this study were relatively narrow, with two key themes emerging: Meeting School Inspection Requirements and Supporting School Improvement. The subsequent sections of this chapter provide a detailed exposition of these themes, as identified through the thematic analysis of the interview data.

5.3.1.1 Meeting School Inspection Requirements

The six interviewees in this study perceived the purpose of school academic evaluation as providing evidence for inspectorates to monitor the educational quality. Two of the policymakers who were interviewed, PM1 and PM2, emphasized the importance of student academic attainment as a critical indicator of a school's academic success. The inspectorates utilized student academic outcome information to judge whether schools had met predetermined goals or standards, considering it as a fundamental element of school inspection in China. PM1 stated:

> *"I believe student academic attainment is the critical indicator of a school's academic success. Based on the goals or standards expected of schools, we use student academic outcome information to examine whether schools have met previously set goals or standards."* (PM1)

PM2 shared the similar perceptions and indicated:

> *"I think the primary purpose of school evaluation is to ensure that schools are meeting certain standards and expectations…… In the context of China, one aspect of school inspection involves using student academic outcome data to assess whether schools have achieved pre-established goals or standards. "* (PM2)

The three headteachers also expressed comparable views regarding the purpose of school evaluation, as exemplified by the following statement:

> *"I think the purpose of school evaluation is to meet one of the school inspection requirements. It serves as a critical component in monitoring high-quality of school education. "* (HT2)

These quotes illustrate that the interviewees perceived the purpose of school evaluation in terms of meeting school inspection requirements. This aligns with the functions of school inspection identified by Scheerens et al. (2003) and OECD (2008), which include assessing the quality of education provided by schools.

It is notable that the purpose of external school accountability was less frequently mentioned by the participants in this study. This may be due to the specific context in China that the primary emphasis of school inspection appears to be on formative evaluation and internal accountability rather than external accountability (Zhen, 2019). This is also reflected in stakeholders' comments in this study. However, the main function of school inspection sometimes involves both formative evaluations aimed at providing advice and summative evaluation focused on holding schools accountable (Scheerens et al., 2003). External accountability for school performance is generally emphasized in many school systems worldwide, including England, the United States, and Australia, where school evaluation results are used to rank schools (Leckie & Prior, 2022). Hence, the question of whether China should emphasize the external accountability function of school inspections emerges as a topic of discussion in the subsequent chapter.

5.3.1.2 Supporting School Improvement

Four interviewees mentioned that school evaluation can serve as an essential tool for enhancing school improvement by identifying areas in need of enhancement. HT1 stated:

> *"I think the purpose of school evaluation is to provide helpful feedback information to schools with identification of their strengths and weaknesses and provide direction for school improvement. For example, if evaluation results show that students' academic performance is poor, the school may focus on the course design or teaching development to address the issue. "*(HT1)

PM2 emphasized that "*evaluation results can help policymakers to monitor policy implementation and establish meaningful goals for schools"*. Another participant, PM1, suggested that "*high-performing schools could share successful experiences with low-performing schools and provide support for their improvement"*. Overall, both policymakers and headteachers emphasized the internal use of evaluation results to support school improvement at both the policy and school levels.

It is evident that both policymakers and headteachers mentioned another widely acknowledged function of school evaluation, as found in the existing literature, which is to facilitate school improvement. Evaluation findings can contribute to data-driven decision-making at the policy level and support self-evaluation at the school level (OECD, 2008).

In summary, while some of the purposes mentioned by participants in this study are in line with those reported in previous literature, others were not emphasized as frequently, such as addressing equity issues in education by focusing on specific student groups, supporting teacher development, and improving the allocation of educational resources. Nevertheless, it is important to note that the multiple purposes of school evaluation highlight the significance of obtaining accurate performance measurements to avoid biased decisions and policies. Therefore, it is crucial to utilize valid and reliable evaluation methods in educational practices. The following section will investigate the effectiveness of the current local junior high school evaluation methods in generating accurate school performance information.

5.3.2 Sub-Question 2.2 *What are interviewee perceptions of the benefits and disadvantages of the unadjusted raw attainment-based school performance measures in the local context of this study?*

In the context of this study, the evaluation of junior high school performance relied primarily on unadjusted raw attainment measures based on the average performance of students in the SHSEE at the end of Grade 9 (age 14/15). The LEA and schools utilized specific indicators, named 'one mark and two rates', to assess performance, including the average marks attained in the SHSEE, the average

pass rate (i.e., the percentage of students scoring 60 or above out of 100), and the average outstanding rate (i.e., the percentage of students achieving 85 or above out of 100) (Du & Hao, 2021). Two themes were identified in this section to demonstrate the interviewees' perspectives on the current unadjusted raw attainment measures of junior high school performance: 'Benefits of Unadjusted Raw Attainment Measures' and 'Disadvantages of Unadjusted Raw Attainment Measures'.

5.3.2.1 Advantages of Unadjusted Raw Attainment Measures

Despite the significant criticism surrounding the use of unadjusted raw attainment measures in school performance, one policymaker and two headteachers identified certain benefits associated with these measures. PM2 stated:

> *"Over the past few years, there have been no changes in the school inspection practice to incorporate the results of the unadjusted raw attainment measures as one of the indicators of junior high school academic performance. I believe that the unadjusted raw attainment measurement results provide a clear and standardized method for assessing school academic performance."* (PM2)

Headteachers tended to highlight the motivational benefits associated with unadjusted raw attainment measures for school improvement. HT1 indicated:

> *"By receiving feedback on our students' performance in the SHSEE, we are able to identify areas where we are doing well and areas that need improvement. This information assists us in making informed decisions and establishing meaningful goals to drive school improvement."* (HT1)

Aanalysis revealed that receiving recognition from the education authority and other schools based on high-performing results provides motivation for some schools to continue their academic success. For example, HT3 revealed:

> *"The recognition we receive from the LEA motivates us to continue our academic success. In other words, high-performing results provide evidence that our teaching is effective in helping students achieve better attainment."* (HT3)

The responses of HT2 demonstrated that motivation for improvement was not only produced between schools, but also within the school.

> *"Last year, our teaching group in the subject of Chinese was rewarded for the outstanding performance of our students by the district education authority. Afterwards, teachers from other subject groups in our school actively learned from our successful experiences. I think this is a good way to stimulate schools and teachers to improve."* (HT2)

In summary, the above statements indicate that both policymakers and headteachers recognize the benefits of the unadjusted raw attainment measures in school performance, including their simplicity, transparency, ability to inform policymaking and goal setting, and usefulness in stimulating improvement. Moreover, interviewees' responses indicated that policymakers, headteachers, and teachers have extensively relied on the unadjusted raw attainment measures for an extended period. Thus, it can be recognized that the unadjusted raw attainment-based evaluation results may aid in facilitating effective communication among stakeholders. However, serious weaknesses of this approach were also highlighted as outlined next.

5.3.2.2 Disadvantages of Unadjusted Raw Attainment Measures

Three interviewees expressed concerns about the limitations of using unadjusted raw attainment-based measures for comparing school performance, particularly in terms of measurement fairness. HT3 compared the performance of his former urban school, where students had higher prior attainment, with his current rural school, where students had lower prior attainment. This comparison highlighted that prior attainment could influence school performance.

> *"I used to be dean of studies in a school in an urban area where the overall students' entrance level (student prior attainment in JHSEE) was higher than in other schools. It is not unreasonable to expect the results of that school to be at average or above average level. I was appointed as headteacher in my current school last year. This school is in a rural area, and the overall student entrance level is lower than average. This year, the measurement of results for our school were at the bottom. However, I know our teachers and students have worked hard. To some extent, this frustrated our teachers. "* (HT3)

Furthermore, two headteachers (HT1, HT2) mentioned that raw attainment measurement results might not accurately reflect the efforts made by teachers and students. This suggests that biased measurement results can negatively impact teachers and students emotionally.

HT2 stated:

> *"It is common to use 'one mark and two rates' as indicators of junior high school academic performance. Although it is reasonable to measure a school's academic performance using the level of student achievement, student academic achievement is influenced by a wide range of factors, including the effort put in by teachers and students, as well as external factors such as home environment. It seems that raw measurements may not always fully capture the impact of these factors and may not always accurately reflect the hard work that teachers and students put into their work."* (HT2)

The responses from HT2 underscored the limitations associated with evaluating schools solely based on unadjusted raw attainment at the culmination of a specific phase of education. Such an approach fails to encompass the broader efforts and contributions of students, teachers, and the school. PM3 shared HT2's perceptions closely and additionally highlighted the drawbacks of neglecting to account for differential effectiveness at the individual level, such as variations between male and female students or disparities between students from rural and urban backgrounds. PM3 indicated:

> *"The 'one mark and two rates' has been used to indicate junior high school's academic performance for a long time. However, we have noticed that these indicators tend to reflect aggregate information at the school level rather than the uniqueness of the school, class and student level. I think it is problematic to talk about school outcomes without caring about individual students. For example, did girls do equally well as boys? Did students from rural areas perform with higher or lower mark than those from urban areas?" (PM3)*

PM3's responses indicate that by focusing only on aggregate information, policymakers may miss important differences in academic achievement that could help to identify areas for improvement and support more targeted interventions to help individual students succeed.

HT1 also pointed out that to achieve a better outcome for schools in performance measures teachers tend to be selective in their attention to certain groups of students to ensure good academic performance. They focused on students who are near the pass standard, implying that they prioritize helping these students meet the minimum academic requirements. They also focused on students with a good chance of achieving high exam marks, indicating that they prioritize students who are likely to achieve excellent academic results. This selective attention to particular groups of students may result in neglecting the needs of students who are neither close to passing nor likely to achieve high marks.

> *"Teachers tend to focus on two groups of students to ensure a good classification performance. One group is students near the pass standard, and the other group is students with a good chance of achieving high exam marks. "* (HT1)

In summary, the thematic analysis reveals various weaknesses associated with unadjusted raw attainment measures. Firstly, it has the limitation to enable a fair comparison. Secondly, it may not accurately capture the effort put in by teachers and students by failing to present the effects of external factors that were outside the school control. Thirdly, it has the limitation to indicating differential effectiveness at the individual level. Finally, it may lead to teachers focusing on specific groups of students, neglecting the needs of other students. These disadvantages suggest that there may be a need to explore new approaches that can facilitate more accurate school evaluations and provide comprehensive information to support school improvement.

5.3.3 Sub-Question 2.3 *What are interviewee perceptions on the concept of VAMs?*

This section examines the interviewees' existing knowledge and understanding of VAMs. Gaining insight into the interviewees' perceptions of VAMs is crucial for the effective implementation of this method, as it can help address any misconceptions or anticipate potential challenges. The analysis yielded three primary themes: "Different Perspectives on the Concept of VAMs," "The Significance of Baseline Assessment," and "The Necessity of Considering Contextual Factors."

The analysis of the interviewees' responses reveals that there were divergent interpretations regarding the meaning and measurement of VAMs. HT1 expressed that "*Calculating the difference between a student's raw attainment in entrance examinations and their performance in school-leaving examinations would provide an indication of a school's added value* "(HT1). On the other hand, HT2 suggested that "*Measuring a school's added value could involve comparing observed student marks with their expected marks in leaving examinations*" (HT2). PM2 defined value-added as "*The school's contribution to student progress in academic attainment*" (PM2). These varied perspectives highlight the different understandings of how VAMs should be conceptualized.

Two interviewees provided their perspectives on the significance of students' baseline assessments when considering VAMs. While HT2 admitted unfamiliarity with the statistical techniques used for calculating expected marks, he acknowledged the connection between student attainment in the JHSEE

and SHSEE. HT2 indicated the following insight based on his experience as a headteacher:

> *"In my experience as a headteacher, I have observed how a solid foundation in one exam can contribute to students' success in future exams. Although I may not be an expert in statistical techniques, I think it is important to understand the relationship between student attainment in the JHSEE and SHSEE."* (HT2)

PM2 shared HT2's understanding and emphasized the following:

> *"In order to accurately assess a school's performance, I think it's essential to take into account students' starting points and the progress they make over time."* (PM2)

Thirdly, both PM2 and PM3 recognized that different schools operating in diverse contexts may face unique challenges and possess varying educational resources. Thus, they emphasized the significance of taking contextual factors into account when utilizing VAMs. For example, PM3 indicated:

> *"I think it's important to consider contextual factors such as school size, status, type, and location. For example, a small rural school may face different challenges compared to a large urban school with a diverse student population. Different types of schools may have varying educational resources. Therefore, I think we need to take into account the context in which the school operates. This may help us to make a fair evaluation of school performance."* (PM3)

In summary, the thematic analysis indicates that while there were different interpretations of VAMs, overall, the interviewees demonstrated a reasonably good understanding of the concept. However, some interviewees exhibited limited technical knowledge concerning the statistical methods employed in calculating VAMs. Moreover, the interviewees highlighted the significance of factors such as baseline assessment and contextual considerations, to some extent, indicating a general understanding of how value-added should be measured. Nevertheless, the varying interpretations underscore the need to clarify the concept of VAMs and establish a shared understanding among stakeholders before implementing VAMs.

5.3.4 Sub-Question 2.4 *To what extent (strong or weak) do interviewees have a motivation to implement VAMs?*

This section aims to examine the level of motivation among interviewees regarding the implementation of VAMs. By gaining a deeper understanding of their motivations, the researcher can identify the

potential challenges and opportunities associated with integrating VAMs into the local school evaluation practice. Two dominant themes have emerged from the data analysis, namely 'Strong Motivation to Implement VAMs' and 'Weak Motivation to Implement VAMs'. These themes will now be elaborated upon in greater detail.

5.3.4.1 Strong Motivation to Implement VAMs

Three policymakers appeared to have a relatively strong motivation to implement VAMs in the evaluation practices of schools. Specifically, PM1 exhibited a strong inclination towards utilizing VAMs to ensure fairer evaluation of school performance and to support their monitoring efforts. PM1 stated:

> *"We usually employ student examination results as one of the means to monitor school performance. Since value-added approaches can enable a fairer school comparison than those based on students' raw examination marks at a single point, we are interested in applying it."* (PM1)

Similarly, PM3 emphasized the efficacy of VAMs in accounting for uncontrollable factors at the school level, which are not addressed in unadjusted raw attainment measures. Although only school-level factors were specifically mentioned by PM3, she acknowledged the limitations of using raw examination results as a measure of school performance. PM3 expressed openness to exploring the implementation of VAMs within the local context, and stated:

> *"We recognize the limitations of the unadjusted raw attainment measures of school academic performance. As I mentioned earlier, such measures fail to consider the influence of contextual factors at the school level. I am open to the idea of VAMs and would like to learn more about how it could be implemented in our context. "* (PM3)

PM2 pointed out that while the unadjusted raw attainment measures may indicate how well students perform in specific examinations and subjects, they may not fully account for the impact of schools in enhancing student outcomes. PM2 indicated:

> *"While unadjusted raw attainment measures may indicate how well students perform in specific examinations and subjects, however I think they may not be sufficient to show the degree to which schools have successfully enhanced student academic performance. From my understanding, VAMs can fulfil this requirement. Thus, I am interested in adopting this evaluation approach."* (PM2)

HT2 also revealed a motivation to implement VAMs. He expressed a desire to explore an approach that provides a deeper understanding of the school's impact on student learning and identifies areas for improvement. HT2 stated:

> *"Although raw examination results were used by the school inspection as one of the indicators of school academic performance, I am still thinking about the approach that can help us better understand our school's impact on student learning and identify areas where we can make improvements. Thus, I am interested to look at how VAMs can provide a comprehensive picture of our school's performance."* (HT2)

Overall, the analysis reveals a considerable motivation among policymakers to implement VAMs in school evaluation practices, with more policymakers expressing strong motivation compared to headteachers. The thematic analysis suggests that interviewees' high motivation stems from their recognition of the benefits associated with VAMs. They perceive VAMs as valuable tools for accurately demonstrating school performance, identifying areas for improvement, and providing comprehensive information on school performance.

5.3.4.2 Weak Motivation to Implement VAMs

Two headteachers expressed reservations regarding the use of VAMs to measure school performance. HT3 indicated that implementing VAMs does not ease the pressure schools face in ensuring that students achieve target marks in high-stakes examinations. Despite acknowledging the potential benefits of VAMs, HT3 stated:

> *"Although I agree that there are potential benefits of VAMs, the reality is that schools are under a lot of pressure to ensure students achieve target marks in high-stakes examinations. Even if we adopt VAMs, the focus remains on examination outcomes, and the pressure to attain high marks in high-stakes examinations is not relieved."* (HT3)

It is important to acknowledge that the issues associated with teaching to the test, commonly associated with the unadjusted raw attainment measures, may also be relevant to VAMs when academic achievement is used as the outcome measure in value-added models. However, further discussion on this matter will be presented in Chapter 6, as it is important to note that outcome measures in value-added models are not solely limited to student academic achievement, as highlighted by Scheerens et al. (2003).

HT1 and HT3 shared a common perception that the decision-making power regarding the implementation of VAMs lies with the LEA, which consequently led to schools paying less attention to reforming their evaluation methods. HT1 stated:

> *"I have no doubt about the potential benefits of using VAMs to evaluate school performance, but the decision-making power for implementing these approaches lies with the local education authority. As a headteacher, I may lack the authority to independently implement such methods, which can limit my motivation to pursue them further."* (HT1)

Similarly, HT3 viewed VAMs more as a tool employed by external inspectors to evaluate the school. Consequently, HT3 expressed less attention and interest in investigating this approach.

> *"I am not sure that VAMs is something that we would use extensively at our school. I see it more as a tool that external inspectors might use to evaluate our school. Therefore, we tend to pay less attention or interests in investigating this approach."* (HT3)

In summary, the thematic analysis indicates that policymakers show higher motivation for implementing VAMs compared to headteachers. Policymakers may have a greater sense of responsibility for implementing new policies, be more aware of the potential benefits of VAMs for school evaluation and improvement and have better access to funding and support. In contrast, headteachers tended to present a relatively low motivation, potentially due to limited decision-making power and perceiving VAMs primarily as external evaluation methods. They may not fully realize that VAMs data can also be used for internal school self-evaluation. It is crucial to acknowledge that although policymakers hold decision-making power, the engagement and commitment of headteachers and teachers are also crucial for the effective implementation and acceptance of VAMs. This point will be further discussed in Chapter 6.

5.3.5 Sub-Question 2.5 *What are interviewee perceptions of the factors that may enhance or hinder the implementation of the value-added approach in junior high school effectiveness evaluation in the context of this study?*

This section investigates the factors that may enhance or hinder the implementation of VAMs, as understanding these perceptions may provide insights into the potential barriers or facilitators to the implementation of VAMs in the local context. The analysis revealed two key themes: 'Factors Supporting the Implementation of VAMs' and 'Factors Hindering the Implementation of VAMs', which reflect the perspectives expressed by the interviewees.

5.3.5.1 Factors Supporting the Implementation of VAMs

Responses from interviewees regarding the conditions that may support using VAMs in school evaluation can be categorized into four aspects. One aspect is associated with political factors. PM3 revealed:

> *"The potential motivation behind the exploration of implementing VAMs at the local level was primarily linked to a new policy document on education evaluation reform. This document drew the attention of provincial education authorities to consider VAMs as a new evaluation method to be utilized within the local context."* (PM3)

PM3 also pointed out that despite having relatively strong motivations to implement VAMs at the district level, they required support from higher-level education authorities to move forward with the plan. It implies that receiving the support from the higher-level education authorities, such as provincial or city level, is perceived as a factor that may support them to implement VAMs.

> *"As grassroots educational management organizations, our capacity to carry out education reform is comparatively limited in comparison to the provincial-level education departments. Therefore, it would be advantageous if higher-level education authorities demonstrate an interest in VAMs."* (PM3)

Another aspect is associated with successful implementation cases. For example, PM1 expressed a positive attitude towards the implementation of VAMs by mentioning the significance of existing cases as a reference for local practice.

> *"We are planning to visit a city in Northeast China to learn about the implementation of the value-added evaluation approach set up in the city to promote school self-evaluation. I think it is good that the value-added evaluation approach has been used to measure school performance in some regions. It provides a valuable reference for our practice in the future."* (PM1)

The fourth aspect is related to the improved data system. Both PM1 and PM2 emphasized the benefits of having an enhanced data system, specifically mentioning how online enrolment and examination registration have streamlined information queries. Additionally, PM2 acknowledged the advantages of having a unique student ID for collecting and matching data requirements for studies. For instance, PM2 stated:

> *"The improvement of the data system enables our students' information to be managed in a more effective way. For example, students' enrolment and examination registration are done online, which facilitates the information query. As I know, the availability of the student unique ID enables you to collect and matching data requirement for your study. "*(PM2)

In summary, the quotes provided above indicate that most interviewees identified several factors that could enhance the implementation of VAMs. They emphasized the significance of government policies, supports from higher-level education authorities, the existence of successful implementation cases, and the presence of improved data systems to facilitate VAMs. However, they also acknowledged potential factors that could hinder the implementation of VAMs.

5.3.5.2 Factors Hindering the Implementation of VAMs

Exploring the factors that hinder the implementation of VAMs is important for identifying potential challenges and developing strategies to overcome them. This enables the adaptation of the evaluation process to the specific requirements of the local context. The analysis yielded five codes that shed light on these hindrances: 'Potential Challenges of Other Education Policies', 'Data Availability and Quality', 'Operational Difficulties', 'Challenges in Interpretation and Understanding of Results', and 'Potential Conflict Regarding the Usage of Unadjusted Raw Attainment Measures Versus VAMs Results'.

Firstly, despite the existence of a policy document related to educational evaluation reform (State Council, 2020), there may be other policies that pose challenges to the implementation of VAMs. For instance, the "Double Reduction Education Policy" (State Council, 2021) was mentioned by PM1 as a potential factor affecting the implementation of VAMs.

> *"The "Double Reduction Education Policy" aims to reduce the burden of excessive homework and private tutoring for students, shifting the focus towards improving classroom teaching and learning. Consequently, there is increased emphasis on formative assessment at the classroom level. This emphasis on formative assessment at the classroom level may impact resource allocation for supporting VAMs, potentially creating challenges for their implementation. "* (PM1)

PM2 and PM3 also highlighted that the government has placed significant emphasis on addressing the long-standing issue of excessive focus on test scores. Against this backdrop, PM2 and PM3 share a common perception that certain individuals may hold a conservative stance towards the implementation of VAMs due to their utilization of academic outcome measures. PM2 expressed this viewpoint:

> *"In recent years, the government has prioritized addressing the problem of excessive emphasis on test scores in education. The government asserts that the objective of basic education is to cultivate well-rounded individuals with a holistic education encompassing moral, intellectual, physical, artistic, and work ethics and skills. Given this context, I think that some individuals may perceive the use of academic outcome measures in VAMs as still focusing on student test scores, potentially deviating from the intended priority of the current situation."* (PM2)

Secondly, interviewees identified factors related to data issues. PM3 and PM2 highlighted the potential consequences of unavailability of student attainment data in the future, specifically concerning the adjustment for student prior attainment due to potential changes in the examination system.

> *"To ease the excessive burden of homework and off-campus tutoring on students, the government has called for a reduction in the number and frequency of large-scale and standards-based examinations during primary and junior high schooling phases. Consequently, the availability of student prior attainment data may become limited in the future."* (PM2)

Even though the desired variables may be accessible for the development of a value-added model, challenges persist in terms of data collection and data quality. HT2 indicated:

> *"Due to political and practical reasons, schools and teachers are prohibited from collecting students' socioeconomic background information to avoid issues of discrimination. Instead, external entities are responsible for gathering this information for various purposes, including education quality monitoring and research projects. These external bodies may encounter difficulties in accessing the required data. "* (HT2)

The way data is collected raises concerns about the reliability and accuracy of data collected by external entities, particularly if parents are unaware or uninformed about the data collection activities. HT3 revealed that "*even though the collected information remains anonymous, parents may still have concerns or complaints regarding these activities*." The acceptance or opposition of parents or students can significantly impact the quality of the collected data and subsequently affect the evaluation outcomes. If parents and students are not adequately informed or do not fully comprehend the data collection activities and have not provided informed consent, they may provide incomplete or inaccurate information, leading to unreliable or biased data. This, in turn, can compromise the overall reliability and validity of the VAMs results. To some extent, this suggests a lack of communication or transparency between external entities and parents regarding the data collection process, as well as the purposes and benefits of data collection.

Thirdly, as discussed in the previous chapter, developing value-added models demands a significant sample size and dataset, along with a specific statistical analysis technique. Therefore, it is unsurprising that thematic analysis highlights the operational difficulties involving limited financial and human resources, and increased workload. For example, PM2 and PM3 indicated:

> *"I think one of the main challenges that we may face in implementing VAMs is a critical shortage of both financial and human resources. There are financial costs associated with conducting VAMs activities to measure school performance, such as employing human resources, costs for necessary training, unexpected costs associated with data collection. We may either included this cost in our extra budget or consider rearranging the budget allocation. However, either way, it will take a longer time for us to make the decision."* (PM2)

> *"One challenge that we may face in implementing VAMs is a lack of human resources with expertise in statistical techniques. I recognize that developing and implementing VANs requires a deep understanding of statistical concepts and methods, and the ability to apply them to educational data. However, we may not have the personnel with the necessary expertise to undertake these activities." (PM3)*

HT1 pointed out that the potential increase in workload could pose challenges for schools and teachers.

> *"The adoption of a new evaluation method would impose an additional workload on schools. However, we are already burdened with various external assessments, inspections, and monitoring, which presents a considerable challenge for us."* (HT1)

Fourthly, the concerns raised by the three headteachers revolve around difficulties in understanding and interpreting VAM results. HT2 expressed the view that*" I think not all school managers and teachers could accurately understand value-added evaluation results, which could result in a minimal impact on school improvement activities."* The other two headteachers shared similar views and pointed out that the statistical nature of value-added approaches and evaluation results would pose obstacles for schools and teachers in effectively utilizing VAM data to enhance student achievement.

> *"We recognize the importance of using data to inform school improvement, but we lack staff who are able to analyse and interpret data accurately and communicate the results in a way that is accessible and meaningful to all staff."* (HT1)

> *"I think it can be quite complicated to use the results for school improvement. While the idea of using data to inform school improvement is certainly appealing, implementing it in practice can be quite difficult. Teachers and school leaders may struggle to understand how to interpret the results and translate them into meaningful instructional changes. These challenges may, in turn, influence the motivation of schools and teachers to fully implement VAMs."* (HT3)

Finally, the theme highlighted following thematic analysis concerns the potential conflicts between the use of results derived from the unadjusted raw attainment measures and VAMs, which were raised by policymakers. PM2 indicated:

> *"No single evaluation information can be a perfect indicator of school performance; therefore, VAMs results can be considered a supplementary indicator of school performance."* (PM2)

PM3 offered a differing viewpoint, advocating for the complete utilization of VAMs results.

> *"I think that VAMs results should be the primary measure of school performance, as they are perceived to be a more robust and reliable school evaluation method. Therefore, the VAMs results can provide a more comprehensive and accurate assessment of the school's impact on student learning and progress over time."* (PM3)

Responses from PM2 and PM3 revealed a fundamental concern regarding the choice between using unadjusted raw attainment measures and VAMs when assessing school performance. In Chapter 6, the implications of these findings will be discussed to provide valuable insights to policymakers when making decisions.

In summary, the feasibility of employing VAMs in the local context is supported by quantitative evidence outlined in Chapter 4. However, the views of policymakers and headteachers indicate that the implementation of VAMs to improve school evaluation and support school improvement cannot be assumed without addressing practical questions. These questions include common issues demonstrated in previous literature, such as a lack of resources, extra workload for schools, and operational problems. Moreover, issues that may be specific in the context of China were also raised, such as the influence of other educational policies, the possibility of the unavailability of prior attainment data, conflicts that may arise associated with the different views on the power of raw examination measures and VAMs in indicating school performance. These questions suggest that quantitative research alone is insufficient to support the implementation of VAMs in local educational practices. More qualitative findings need to be combined with the quantitative results to provide sufficient evidence for policy decision-making (Feng & Zhou, 2022; Peng & Zhang, 2021).

5.4 Chapter Conclusion

In conclusion, the qualitative findings in this chapter provide additional insights into the perceptions of local policymakers and headteachers regarding the benefit, disadvantages, and the potential of employing VAMs in school performance evaluation practices in the sample district. It contributes to a more comprehensive understanding of the issues and challenges related to the implementation of VAMs in school performance evaluation practices in the local context. It also provides additional evidence for policymakers on decision-making in bringing VAMs into local educational practice. The

following chapter will discuss the findings in this chapter together with findings from Chapter 4 and other chapters to address the overall research aim.

Chapter 6 Discussion

6.1 Chapter Introduction

This chapter aims to interpret and discuss the key quantitative and qualitative findings presented in Chapters 4 and 5, within the context of the previous research outlined in the introduction and literature review chapters. Section 6.2 will examine the key quantitative findings, while section 6.3 will focus on the key qualitative findings. The discussions presented in this chapter will provide insights overall and that will help to answer Research Questions 1 and 2, as well as the related sub-research questions.

6.2 Discussion of the Key Quantitative Findings

RQ1: What is the range and extent of school and class academic performance in 11 public junior high schools in the W district in Southwest China based on Raw, Value-added (controlling for prior attainment only), Contextual value-added (controlling for not only prior attainment but also student background characteristics, class, and school context) measures?

To address Research Question 1 and its sub-questions 1.1-1.6, Chapter 4 presented and analysed the findings of four MLMs reviewed in Chapters 2 and developed in Chapter 3: the Raw, VA, CVA-1, and CVA-2 models. These models differ in their use of explanatory variables adjusted for at different levels. The section below will provide a discussion of the key quantitative findings.

6.2.1 Changes in the Range of School Academic Performance Highlights the Significance of Implementing VAMs in Educational Practice to Provide a Fairer Comparison of School Performance

Consistent with previous studies (e.g., Leckie & Goldstein, 2017; Munoz-Chereau & Thomas, 2016; Thomas, 1998), this study revealed a significant change in the range of raw and VA academic performance among the 11 schools examined. In terms of SHSEE total scores, the arithmetic means across schools ranged from -0.86 to 0.74, while this range was substantially reduced to points from -0.37 to 0.36 in the value-added results estimated by the CVA-2 model (which ranged from points above the average to points below the average). Significant reductions in the ranges were also observed for SHSEE Chinese, Mathematics, and English. These findings can be attributed to the adjustment for

school intake differences, which accounted for a significant proportion of the initial variation and brought considerable reduction in school-level variation.

The Raw model showed that the proportion of variance attributable to the school level (school effect) ranged from 12.73% to 15.37% across four outcome variables (Total, Chinese, Mathematics, English). The proportion of variance attributable to the class level (class effect) was somewhat higher and ranged from 20% to 27.05%. However, when considering the CVA-2 models that incorporated student prior attainment, background characteristics, and the class contextual factor, the percentage of variance across the four student academic outcomes attributable to between-school differences reduced to a range of 8.22% to 14.36%. Similarly, the corresponding figures for between-class differences ranged from 6.79% to 8.16%. Comparison with the findings of IEEQC project, which offered more rigorous estimates of school and class effects discussed in Chapter 2, indicated slightly higher school effects and slightly lower class effects in this study compared to the IEEQC project. These findings align with Thomas's (2020) suggestion that school effects in different regions of China revealed variation and tend to be slightly larger than equivalent results in the UK.

While it is important to acknowledge the limitations of presenting changes in school rankings, as discussed in Chapter 4, it is still valuable in illustrating the practical significance of employing VAMs to obtain accurate performance measures. For example, when examining the outcome measure of SHSEE total scores, two out of the eleven schools were initially ranked towards the lower end based on raw standardized mean total scores. However, once student prior attainment and contextual factors were considered, their rank positions increased dramatically to near the top. Almost half of the schools observed substantial changes in their rank positions, moving up or down by more than two ranks, when transitioning from the raw scores to the value-added results. The findings highlight the substantial impact that the implementation of VAMs can have on school rankings and the importance of utilizing such approaches for a fairer evaluation of school performance.

In short, while student raw attainment measures can highlight attainment gaps between schools, they fail to account for school intake differences (OECD, 2008). As a result, raw attainment measures may unfairly attribute sole responsibility for student outcomes to schools. Moreover, raw attainment measures may be biased in favour of schools with high-achieving intakes, leading to frustration among

schools that demonstrate significant student progress compared to others. In contrast, VAMs provide a more accurate measurement of school performance by isolating the contribution of schools to student academic progress. By considering factors such as student prior attainment and contextual variables, it is argued that VAMs generate more robust results and offer a fairer comparison of school performance compared to raw attainment measures. This argument aligns with previous studies conducted in Western countries and China (e.g., Munoz-Chereau, 2013; Munoz-Chereau & Thomas, 2016; Peng et al., 2006; Salim, 2011; Thomas, 2020; Zhang, 2016), which emphasize the need for a range of valid information, including information obtained from VAMs, to judge a school's performance.

6.2.2 In Terms of the 'Goodness of Fit' of Models, More Complex Multilevel Models (CVA-2) are Suggested to Analysis School and Class Academic Performance

The quantitative findings of the study emphasize the importance of including the class level in value-added models. This is demonstrated by the better fit of data from three-level Raw models, which resulted in an approximate 22% reduction in the between-school variation. This provides new evidence in the context of China that the extent and significance of school residuals can vary depending on the levels controlled for in the analysis (Munoz-Chereau & Thomas, 2016). Therefore, it is meaningful to examine the fit of a variety of models to determine the appropriate number of levels to include (Kyriakides & Charalambous, 2004). Additionally, recognizing the influence of class grouping on student outcomes is substantively important because teaching and learning or teacher effectiveness is essential in improving student outcomes. Overall, the study highlights the importance of considering multiple levels of analysis in educational research to obtain a more accurate understanding of the factors that contribute to student outcomes.

As noted in previous literature (e.g., Leckie & Goldstein, 2019; Leckie & Prior, 2022; Thomas, 2001; Thomas & Mortimore, 1996), the modelling approach used to measure school effectiveness can significantly impact the estimation of school and class effects. However, in the Chinese context, there has been limited discussion and comparison of value-added models (Yang & Zhang, 2022). Therefore, the results of this study, which evaluated four different models based on their goodness of fit, provide important evidence for choosing the most appropriate model. Study results align with the widely

accepted view that unadjusted raw models do not adequately isolate the contributions of schools or classes, as they fail to account for initial differences between schools (Leckie & Prior, 2022). When student prior attainment was added as an adjustment in the value-added model, the percentage of total variance explained across four outcome variables increased to an average of 45.44%. The average school-level variance explained was 50.41%, and the average class-level variance explained was 67.08%. Despite the relatively small sample size, prior attainment proved to be a powerful predictor of school and class contributions to student outcomes. These results reinforce the importance of adjusting for prior attainment in evaluations of school performance in the Chinese context, as discussed in Section 2.4.2.1 of Chapter 2. Accounting for prior attainment is necessary to provide a valid and fair assessment of schools' performance, as emphasized in previous research (Zhang, 2016).

Moreover, two alternative models, CVA-1 and CVA-2, were developed to enhance the explanatory power and goodness of fit compared to the VA model. The CVA-1 model controlled for additional student sociodemographic variables, while the CVA-2 model controlled for additional class-level contextual variables based on CVA-1. The results showed that both CVA-1 and CVA-2 models had better goodness of fit than the VA model. However, the CVA-2 model, which controlled for student prior attainment, student background variables, and the class prior mean attainment, demonstrated the highest goodness of fit among the three models. Therefore, the CVA-2 model is considered the optimal model to fit the data collected in this study and to estimate the range in school and class value-added effects.

However, it is important to note that there is less consensus in findings regarding the need to adjust for school and class composition. Some studies have indicated that adjusting for school compositional variables has minimal impact (Leckie & Prior, 2022; Thomas & Mortimore, 1996; Timmermans et al., 2011). Moreover, there may be a relationship between school policies and practices and school composition. Consequently, incorporating school and class composition as control variables might eliminate some of the combined effects of school composition and school policies and practices, potentially leading to an underestimation of the real differences in effectiveness between schools (Thrupp et al., 2002; Willms, 1992). As noted by Timmermans and Thomas (2015), careful consideration was needed to decide whether the school context should be included in value-added

modelling. Therefore, it is important to test the statistical significance of individual explanatory variables when developing value-added models and comparing and reflecting on possible explanations of the estimated results generated by different models to make an appropriate model choice.

6.2.3 Application of Models Should Consider the Local Context and Changing Circumstance

The above discussion of key findings underlines the necessity of developing complex multilevel models for analysing school academic performance in seeking to make like-with-like comparisons. However, education systems may make different choices of models based on the practical conditions or policymakers' preference (Leckie & Prior, 2022). Firstly, if a school system wants to employ the optimal model defined in literature, it may face the issue of data availability and quality. For example, disruptions caused by the COVID-19 pandemic may lead education systems to change their choice of model due to data availability. In the context of this study, the COVID-19 pandemic resulted in a delay in the administration of the end-of-phase tests (SHSEE) and posed significant challenges to data collection, as the routine management process of education authorities and schools was disrupted. Taking the UK as an example, the government chose not to calculate Progress8 for 2020 and 2021 because of the cancellation of the GCSE examinations (Leckie & Prior, 2022).

Secondly, given the complex statistical findings, it is important to consider the end user's preferences in choosing a model for analysing school academic performance. While a complex model may be regarded as optimal from a technical perspective, end users may favour a simpler model that is easier to comprehend (Leckie & Prior, 2022). Thus, there is a need for a balance between statistical rigour and practical utility when selecting an appropriate model.

In addition, the choice of model may also be influenced by new orientations of education reform in a particular context. For instance, in China, the current education reforms prioritize the development of students' core competencies and values, emphasizing the cultivation of "well-rounded" students (Xin, 2019). Therefore, using VAMs that solely rely on academic achievement to evaluate school performance may face criticism for overlooking non-academic outcomes. This concern has also been raised in the UK (Prior et al., 2021). As a result, more research is needed to explore how VAMs can be extended to include non-academic subjects or non-cognitive outcomes, to provide a more comprehensive judgment of school performance.

In conclusion, while MLM has been shown to improve model fit and measure school and class effects more accurately, the practical choices of education systems may differ due to various factors such as local context and changing circumstances. Thus, it is important to consider local context (e.g., users' preference, data availability, data quality) and changing circumstances (e.g., uncertain influence brought by the COVID-19 pandemic on estimates of school performance, the impact of new education policy on the reform of examinations).

6.2.4 Using VAMs for School Improvement Should be Recognized

Although VAMs have been used to rank and compare schools in many educational systems (Munoz-Chereau et al., 2020), their implementation extends beyond merely measuring school performance and is associated with linking school measurement to school improvement (Thomas, 1998). The results of this study suggest that value-added data can provide comprehensive information that can be used for self-evaluation and to promote school improvement. A further discussion of the relevance of the information provided by VAMs for school improvement is provided below.

6.2.4.1 Results of Differential School and Class Effects Can Support Making Subject-Specific Strategies

Firstly, VAMs can provide detailed information not only on overall school performance, but also on the performance of individual subject departments. In this study, the CVA-2 model revealed that the percentages of remaining variation attributable to schools were 8.22%, 12.29%, and 10.25% for Chinese, Mathematics, and English, respectively. These findings suggest that schools have a greater influence on Mathematics learning than on the other two subjects. Additionally, the percentages of remaining variation attributable to classes for the three subjects were also analysed, with a more significant influence from classes found in English learning.

Regarding the differential school effects across Chinese, Mathematics, and English curriculum subjects, the correlations between any two subjects were observed to range from 0.73 to 0.80. These subject correlations are higher compared to the findings from IEEQC project that equivalent figures were ranged from 0.57 to 0.89 (Thomas, 2020). They are also higher than what might be anticipated based on comparable findings in the UK (e.g., Munoz-Chereau & Thomas, 2016; Sammons et al.,

1996). The relatively higher subject correlations found in this study indicate that schools that perform well in one subject are likely to perform well in the other subjects as well. The difference between findings in this study and those in IEEQC project and some studies in the UK suggests that the relationships between subjects within schools may vary across different regions in China or countries.

On the other hand, differential class effects were found to be relatively low across the three curriculum subjects (correlations between any of two subjects ranged from 0.41 to 0.54). This suggests that classes that perform well in one subject may not be guaranteed to perform well in the other subjects. It may indicate that there may be potential variation in teaching methods, teacher expertise, and classroom activities within classes across different subjects. In addition, the low correlation of class effects between subjects may suggest potential variation in student characteristics. For example, it is possible that certain classes may be assigned students who have higher scores in specific subject in entrance examinations or possess other favourable academic attributes. These variations in student characteristics may potentially influence their performance in different subjects.

In short, schools can utilize above information to conduct further analysis of the underlying reasons for differences in subject results within schools. For example, such differences may result from variations in school policies, development planning, or the quality of teaching in different departments (Thomas & Mortimore, 1996). Moreover, this information can provide the evidence basis for supporting school improvement by considering subject-specific teaching strategies, address individual student needs, and potentially leverage teacher expertise in specific subjects to optimize student outcomes across the curriculum.

6.2.4.2 Results of the Relationship between Explanatory Variables and Student Outcomes Can Provide Information Guiding School Improvement Efforts

As noted in Chapter 4, isolating the contribution of schools requires testing the relationship between student achievement and variables outside the control of the school that need to be adjusted. This information can be useful for policymakers and school managers seeking to understand the extent of the relationship between those variables and student outcomes. For example, private academic tutoring is rarely used in the value-added relative studies in China (Guo & Wang, 2021). However, it is notable that the results of the VA, CVA-1, and CVA-2 models in this study show for the first time using MLM

that private academic tutoring has a statistically significant positive effect on student SHSEE outcomes, particularly in English. This finding provides feedback to schools that students' higher academic achievement may be partly due to private academic tutoring and suggests a potential inequity in English teaching provision across the sample schools.

Moreover, it is notable that this study includes all prior attainment variables (JHSEE Chinese, Mathematics, and English), which is not commonly used in other studies in China. The findings indicate that the correlation between JHSEE Mathematics scores and SHSEE Total scores is relatively higher than in JHSEE Chinese and English. This finding is in line with the finding of IEEQC project that students' prior proficiency in Mathematics may have a greater impact on their overall academic achievement across subjects and seems to play a more a more influential role in students' overall performance compared to other two subjects. This information appears to support school improvement decisions such as the potential need to consider allocating resources and interventions accordingly to address the specific needs of students in Mathematics or may consider strengthening the teaching and learning strategies in Mathematics to enhance students' foundational skills.

6.2.4.3 Findings of the School Effects on Gender Groups Can Provide Valuable Insights in Individual Students

Value-added results can provide informative insights into groups of students. For instance, this study found statistically significant and negative effects of student gender (boys in comparison to girls) on student SHSEE outcomes in Chinese and English, indicating that girls achieved significantly higher progress scores than boys in these subjects. This finding is consistent with prior research conducted in China (e.g., Fan & Gao, 2019; Lv, 2015; Shao et al., 2021; Wang et al., 2009) and in Western countries (Thomas & Mortimore, 1996). However, a reverse finding was observed in research conducted in sub-Saharan Africa, where girls achieved significantly lower scores than boys (Salim, 2011). Interestingly, this study found that student gender was statistically insignificant to student SHSEE outcomes in Mathematics, whereas prior research in Chinese literature reported a reverse finding (e.g., Fan & Gao, 2019; Shao et al., 2021; Wang et al., 2009). Overall, value-added results can provide valuable evidence for schools to pay attention to issues related to individual student characteristics, such as gender equity and differential learning experiences (Salim, 2011; Sammons et al., 1996; Thomas et al., 1997). Further

analysis of these findings could provide more profound information to investigate the sources of inequity in student groups, with the aim of improving the progress of all students.

Based on the above discussion, it is evident that value-added data can provide detailed information on various subjects, specific explanatory variables, and student groups. By providing comprehensive and sophisticated information, VAMs can assist school staff in considering and exploring critical factors that may explain variations in a school's academic performance. As a result, VAMs have the advantage of informing policy development and school management in various areas (OECD, 2008). Therefore, although there are caveats in using quantitative measures both raw and value added (e.g., Foley & Goldstein, 2012; Goldstein, 2011; Goldstein & Spiegelhalter, 1996), it is essential to recognize the shift from using value-added data for school measurement to using it for school improvement and improving equity, as emphasized by Thomas (1998).

6.3 Discussion of the Key Qualitative Findings

RQ2: What are stakeholders' perceptions of the potential benefits, disadvantages, and implementation of the value-added approach in school evaluation practices when local contexts are considered?

To explore RQ2 and sub-questions 2.1-2.6, this study employed a qualitative research approach by conducting semi-structured interviews with a small group of participants consisting of three local policymakers from the district education authority and three headteachers from public junior high schools. The aim of the interviews was to gather participants' insights and perceptions related to the purpose of school evaluation, the concept of value-added, the motivations of implementing VAMs, and the factors that may facilitate or hinder the implementation of VAMs in their respective schools. The key findings from this inform the discussion below.

6.3.1 Implementing VAMs to Enhance School Self-evaluation

The participants' perceptions of the purpose of school evaluation were primarily associated with school inspections and improvement. Other policy objectives, such as accountability, parental satisfaction, school management, and educational equity, were mentioned less frequently. This finding may be explained by the local political and cultural context, particularly given that the study was conducted in a disadvantaged region, and the education management organization at the district level is a relatively

grassroots organization with limited resources that can be actively allocated. On the other hand, while it is common for school systems worldwide to adopt school accountability systems and publish school performance measurement results online (Leckie & Prior, 2022; OECD, 2008), China does not have a public accountability system. Therefore, participants in this study may not have been motivated to mention it.

In the context of China, social and cultural factors may contribute to the lack of mechanisms for public accountability. While the publication of schools' value-added results in some Western countries is intended to promote school choice, this is not encouraged during junior high schooling in China due to concerns that families with higher socio-economic status may have an advantage in choosing better schools (Zhen, 2019). Moreover, education equity rather than competition is emphasized by the Chinese government in the compulsory education phase that includes junior high schooling. Additionally, the publication of school performance information may create a culture of blame, failure, and labelling (Xie & Zhang, 2021; Yang et al., 1999). Thus, the Chinese government may not want to publicise the possibly wide disparity in school results and inequity across regions.

Given the limited emphasis on public accountability in China, the implementation of VAMs should place particular emphasis on enhancing school self-evaluation to enhance educational outcomes (Elliot et al., 1998). VAMs could provide detailed and comprehensive feedback (as discussed in Section 6.2.4) and serve as valuable evidence for school self-evaluation. Specifically, VAMs results could identify teachers, programs, or interventions that have a significant positive impact on student learning. This information appears to guide school self-evaluation by highlighting successful strategies that can be replicated and developed within the school to improve overall student outcomes. The VAMs feedback can also assists school leaders in facilitating improvements by recognizing progress throughout the entire institution and identifying performance that exceeds or falls short of expectations across all curriculum domains. In addition, VAMs enable a fair measure of teacher effectiveness. Thus, rather than relying solely on the raw examination results, VAMs results may support the school to ask the right questions about teaching and learning practices within schools (Teddlie & Liu, 2008).

6.3.2 Both Policymakers and Headteachers Play an Important Role in Using VAMs for School Improvement

Findings suggest that policymakers in this study are more motivated than headteachers to incorporate VAMs into local practices. This is likely because the policy to explore VAMs is being advocated at the national level, with "high-quality development" in education being a significant priority (State Council, 2020; Xin & Li, 2020). To achieve this goal, a series of education reform policies have been introduced, with a focus on improving the validity and reliability of school evaluations. The limitations of using raw attainment as the sole measure of school performance have also been acknowledged (Gao & Song, 2021). As discussed in Chapter 5, a recent policy document relevant to education evaluation in China (State Council, 2020) advocates for exploring advanced evaluation methods to accurately measure school performance, identify best practices, and highlight areas for improvement (Wang & Pai, 2022). Increasing attention by local policymakers can also be seen from the currently increasing practical cases in some regions, such as Liaoning Province in northeast China, mentioned by the participants. This was publicized at a sub-forum called "the Exploration and Practice Innovation of Value-added Evaluation" at the China Basic Education Forum (CBEF) held recently in 2023. Experiences of policymakers in Liaoning Province gained from the district level to the province level have been shared. Broadly speaking, the central proposition of this forum appears to be that within the current context of China, there is a need for comprehensive exploration and practical implementation of VAMs. This exploration may primarily involve the establishment of a conceptual framework for VAMs and the investigation of the potential breakthroughs this methodology can yield in supporting school improvement.

However, although the policymakers in this study support the implementation of VAMs in educational practice, they do not perceive it as an urgent measure. It is possible that their responses reflect a lack of comprehensive understanding of VAMs. For example, policymakers in this study regarded student academic attainment as the only outcome that could be used in VAMs. Therefore, as discussed in earlier sections, policymakers may consider postponing the agenda of implementing VAMs under the requirement of correcting the bias of overemphasizing academic results in examinations. Despite this, it is still necessary to emphasize that VAMs can provide relatively accurate performance information

that can serve as a basis for raising questions and possible action within schools and classes that supports continuous improvement, as discussed in Section 6.2.4. In other words, using scientific and reliable value-added data as the basis for internal policy decision-making to support improvement is crucial. Once value-added information is obtained, policymakers can better analyse how to allocate educational resources, implement appropriate programs to improve student performance, and set meaningful goals, as emphasized by the OECD (2008) and Wen & Sun (2022).

The responses of two headteachers in this study indicate a slightly passive attitude towards promoting the implementation of VAMs. They view their role as policy implementers and the school as the object being evaluated, with the VAMs results serving as a potential indicator of whether schools are meeting the standards set by the authorities. However, this perspective overlooks the potential for VAMs to support self-evaluation and school improvement. One of the main purposes of developing VAMs is to have a positive impact at the school level and increase the performance of schools (OECD, 2008). This impact is dependent upon schools' ability to effectively interpret and act on value-added information (Saunders, 2000). For example, school self-evaluation could be enhanced by analysing value-added scores for specific groups of students, which can indicate how different student groups within the school/department have performed (OECD, 2008). Additionally, value-added data can serve as diagnostic assistance for identifying barriers to students' educational progress, raising questions about teaching and learning, and guiding improvements (Zhen & Song, 2021). Therefore, it is crucial to recognize the use of VAMs for school improvement in the context of China. Headteachers should be encouraged to adopt a more motivative attitude towards the implementation of VAMs and explore how value-added data can support self-evaluation and improvement in their schools. This approach can help to ensure that VAMs are utilized effectively to support positive changes in school education systems in China.

In short, monitoring school performance alone is insufficient for improving performance. To enhance school effectiveness, it is essential to investigate the relationship between school performance measurement results and the conditions that appear to enhance or hinder school effectiveness in a specific school (Thomas, 1998). Therefore, a school is not only an evaluation object defined by headteachers, but also a unit of action. Information obtained from VAMs can be utilised for a variety

of school improvement purposes, but only if actors such as teachers and headteachers, who can influence the process and/or outcomes, utilise that information (OECD, 2008). Thus, regardless of the dominant role played by policymakers in promoting the implementation of VAMs, enhancing the knowledge base of headteachers about VAMs and their ability to interpret, raise questions and act upon value-added evaluation information is of equal importance in achieving successful implementation.

6.3.3 Although Some Factors Support the Implementation of VAMs, the Road Towards Including VAMs in Local Educational Practices is Just Starting

Participants in this study identified several factors that may facilitate the implementation of VAMs. Firstly, as noted above, political reasons may support the adoption of VAMs in China, as policymakers are influenced by national educational evaluation policy that advocates for exploring value-added evaluation. Additionally, fairness is viewed as a critical factor in the evaluation of schools in China (Guo & Wang, 2021; Peng et al., 2006). Methodological considerations were also raised as important factors that could enable the successful implementation of VAMs. Specifically, the increasing use of MLM in VAMs research in China provides support for enhancing statistical techniques for measuring value added. The development of information technology infrastructure, such as a student database that includes unique student identification and scores in the JHSEE and SHSEE, could also facilitate the collection and use of data for value-added modelling.

Despite the factors that support the implementation of VAMs, the existence of numerous barriers to implementing VAMs in school evaluation practices were identified by the participants. Some barriers are in line with the difficulties highlighted by Chinese scholars (e.g., Feng & Zhou, 2022; Ma, 2020; Ren, 2022; Xin, 2020; Zhen & Song, 2021), including inadequate availability and quality of data necessary for conducting value-added evaluation, the undesirable side effects of test-based outcome measures, insufficient human resources with expertise in statistical techniques, extra burden on schools and teachers, and difficulty in comprehending value-added methodology. On the other hand, the study revealed additional challenges not frequently mentioned in Chinese literature. For example, the education authority at the district level has limited influence in making policy decisions concerning the implementation of new school evaluation methods without support from the education authority at the city level.

Moreover, participants raised the issue of whether VAMs should be considered the only indicator of school performance, although this issue was less discussed in the Chinese literature reviewed in Chapter 2 (Feng & Zhou, 2022; Ma, 2020; Xin, 2020). This study recognizes the limitations of VAMs and does not suggest that they should be the sole indicators used to measure school performance or used as the exclusive basis for high-stakes decisions. Instead, value-added indicators should be used alongside other performance indicators to raise questions and provide a more comprehensive understanding of a school's performance. It is also important to note that VAMs are estimates, not a perfect measure and are based on past performance and may not accurately predict a school's future performance or provide guidance for current students. Therefore, policymakers and educators should use VAMs in combination with other measures to gain a comprehensive understanding of school performance and make informed decisions.

6.4 Chapter Conclusion

This chapter has provided a comprehensive interpretation and discussion of the key quantitative and qualitative findings presented in previous chapters. Through new analyses of the range and extent of school and class academic performance, it has been demonstrated that VAMs can provide a fairer comparison of school performance, with more complex models suggested for analysis, in line with previous research. Furthermore, new evidence on the importance of considering local context and changing circumstances when applying VAMs has been highlighted, alongside the recognition of the role of VAMs in school improvement. The stakeholders' perceptions of the potential to implement VAMs in school evaluation practices have been explored, with the need to enhance stakeholders' understanding and knowledge of the concept and methodology of VAMs in education. The qualitative findings also suggest that while there are some factors supporting the implementation of VAMs, there is still a long road ahead in implementing VAMs in local educational practices. The final chapter will discuss what is new about this study and findings, the implications of this study in policy practice, the limitations of the research methods, and suggestions for future research.

Chapter 7 Conclusions

7.1 Chapter Introduction

This study aims to examine the range and extent of school and class academic performance in public junior high schools when VAMs are used and to explore stakeholder perspectives regarding the potential benefits, disadvantages, and implementation of VAMs to support school evaluation and improvement when the local context is considered. This chapter will conclude the thesis by clarifying contributions, providing theoretical and practical implications, reflecting on the limitations, and end with suggestions for future research.

7.2 Unique Contributions of the Current Research

This study has contributed significantly to the theoretical and methodological knowledge relating to VAMs and their implementation in the Chinese context, specifically in relation to the evaluation of public junior high schools.

First, the study utilized Scheerens' integrated model of school effectiveness (Scheerens, 1992) as a valuable conceptual framework for analysing the local educational landscape in China. This theoretical framework allowed for a comprehensive examination of the factors that contribute to school effectiveness, including the school's organizational structure, teaching and learning practices, and student characteristics.

Secondly, this study contributed to the methodological advancement of VAMs research in school evaluation through the application of a mixed-methods approach. Compared to most relevant Chinese studies that predominantly employed quantitative methods, this study provided richer evidence and a deeper understanding of VAMs and their implementation in the Chinese context (Du & Yang, 2011; Hu et al., 2022; Lv, 2015; Shao et al., 2021; see Chapter 2 for more details). While there were studies that involved Chinese scholars' discussions on the implementation of VAMs, this study added to the literature by including the perspectives of practitioners such as local policymakers and headteachers, thus providing valuable insights into bringing VAMs into practice and perceptions of implementing VAMs.

Thirdly, this study employed multilevel models with increased complexity, incorporating schools, classes, and student levels. This approach represented a significant improvement over previous value-added evaluation studies in China (e.g., Fan & Gao, 2019; Peng et al.,2013; Shao et al., 2021), which typically focused solely on school and student levels. The results revealed that CVA-2 models demonstrated the best fit for explaining the total variance, while still demonstrating significant differences between schools and classes within schools. These findings suggested that measuring school performance in the local context of this study required the use of complex modelling approaches to produce reliable and valid results. However, as discussed earlier, the selection of appropriate models was not straightforward, particularly when the resulting school effects varied considerably across different models. Therefore, this study recommended that policymakers considered the purpose and practical conditions, such as data availability and quality, the preferences of end-users and changing circumstances, unexpected events, and new policy orientations, when selecting value-added models.

Furthermore, this study offered valuable insights into the factors controlled in the models. As expected, not all variance in students' SHSEE outcomes could be accounted for, as some unmeasured factors beyond the school's control might have played a role. Despite that, this study firstly employed several prior attainment variables to provide a more detailed understanding of the factors affecting student performance. Results indicated that the inclusion of all prior attainment variables explained the highest proportion of total variance in student performance. Unlike most previous Chinese studies that often employed outcome measures in one, two, or sometimes three subjects without providing a clear explanation or rationale for the chosen subjects (e.g., Fan & Gao, 2019; Shao et al., 2021), this study highlighted the importance of employing various types of prior measures related to the curriculum and provided statistical evidence to support this assertion. Moreover, this study employed private academic tutoring as an explanatory variable and indicated a statistically significant positive effect on student SHSEE outcomes for almost the first time. Additionally, by controlling for the class mean prior total attainment in the CVA-2 models, this study demonstrated a significant reduction in the percentage of variance in student SHSEE outcomes attributed to class differences. These new findings were likely to generate increased interest among researchers and educational authorities, prompting them to investigate the underlying reasons behind these outcomes.

In addition, this study contributed new evidence to the field of research on VAMs. Specifically, unlike the ITDEQC and IEEQC projects (Thomas, 2005; 2020; Thomas et al., 2011; 2012; 2015) that focused on senior high schools in the regions of East and West, this study revealed equivalent new findings for junior high schools in the Southwest region. By incorporating the class level in the multilevel models, this study demonstrated that class-level differences significantly impacted junior high students' progress in Chinese, Mathematics, and English, highlighting the importance of specifying the class level when estimating school effects. The study presented additional evidence of differential school effects across three different curriculum subjects in junior high schools, supporting the suggestion made by the ITDEQC and IEEQC projects that whole school policies implemented across subject departments tended to be influential (Thomas, 2020). By utilizing mixed methods, this study provided qualitative findings in Chapter 5 that contributed to understanding the implementation of VAMs in the local context of China. Through in-depth interviews with policymakers and headteachers, the study identified several factors that potentially hindered the implementation of VAMs, including data availability and quality, model complexity, influence of external stakeholders, and limited use of results in school improvement. These findings suggested that the implementation of VAMs in local school evaluation practices and school self-evaluation took time and needed support. Policymakers and headteachers needed to carefully consider unintended consequences and collaborate to improve the methodology, generate more reliable estimates, adapt to changing circumstances, and enhance the capacity of practitioners to use measurement results. The discussion also highlighted the tensions surrounding calls for increasing public accountability of schools in China and the use of VAMs for this purpose.

7.3 Implications for Educational Practices

The limitations of relying solely on raw attainment measures to evaluate student performance were extensively discussed in Chapters 2 and 4 of this thesis. Despite the relatively small scale of this study, the findings suggested that VAMs represented a valid and effective tool for measuring school performance in Chinese junior high schools. However, it is important to note that the successful implementation of VAMs depended not only on the advanced methodological underpinnings demonstrated in this research but also on the careful consideration of a number of key issues raised

throughout the thesis. Accordingly, the implications of this study were mainly related to the issues discussed in previous chapters and offered important insights for educators and policymakers considering the implementation of VAMs in local educational practice.

Firstly, a top-down approach to VAMs implementation might have been more effective in the local context. As discussed in Chapter 6, while policymakers at the district level may have had the motivation to implement VAMs, receiving support and guidance from the education authority at the city level would have been a significant boost to promoting their implementation.

Secondly, implementing VAMs took time, and initial motivation may have naturally weakened over time. Therefore, it was essential to build capacity for policymakers and headteachers to understand the range of potential uses of VAMs information and to obtain a long-term commitment from all stakeholders involved, including researchers, policymakers, headteachers, teachers, and others. This commitment was necessary because the implementation of VAMs did not motivate change by itself, but raising questions and acting on the value-added information could bring about change. For example, using value-added data to support data-based decision-making could empower policymakers and school managers to analyse variation in school performance, identify areas of best practice, and areas that needed improvement. However, it could not be taken for granted that policymakers and headteachers had the capacity to interpret the information and know what to do with the data. Therefore, appropriate training was necessary before VAMs could be used in local educational practice, in line with the arguments of previous research (Thomas, 2005; 2020; Thomas & Peng, 2011; Thomas et al., 2012; 2015).

Although this issue was addressed in earlier sections, it is important to reiterate that school evaluation methods should have been integrated into the existing school evaluation system (OECD, 2008). In the Chinese context, school evaluation was conducted through school inspectorates, and the evaluation results were not publicly disclosed. Therefore, instead of considering the feasibility of publishing schools' value-added results and in what manner, it might have been more effective to strengthen the internal and external accountability of school inspectorates and enhance their capacity to utilize VAMs in school evaluation.

Additionally, it is important to note that although VAMs provided a comprehensive measure of school

performance, they should not have been considered as the sole indicator of school quality. Other factors such as student engagement, school climate, teacher quality, and parental involvement also contributed to school performance (Leckie & Prior, 2022). Therefore, using VAMs alongside other measures would have provided a more accurate and comprehensive picture of school performance. In terms of the different value-added models used in the study, this study suggested that presenting both VA and CVA model results could have helped stakeholders understand a more comprehensive picture of school performance and identify potential areas of improvement. This study also highlighted the importance of choosing appropriate covariates for the CVA model, as this could have significantly affected the results. Hence, presenting both VA and CVA model results could have provided valuable information for school evaluation and improvement, but the appropriate interpretation and use of these results would have required careful consideration of the caveats, context, and stakeholders involved.

Finally, pilot programs were necessary to ensure the effective implementation of VAMs. As this study was limited to one district, further empirical research based on the local context was needed. However, the objective of pilot programs should not have been limited to merely testing the value-added model in the local context. It was also crucial to explore operational and implementation issues, including decisions on the choice of the value-added model, to assess how value-added information, in conjunction with other indicators, could provide a comprehensive picture of school performance (OECD, 2008). A pilot program could have also provided an excellent opportunity to develop an engagement and communication strategy for stakeholders, addressing the issue of developing a long-term commitment from them.

In short, the findings of this study provided useful implications for local policymakers in exploring the feasibility of applying VAMs in local educational practice. However, it was important to note that the transition from research findings to a practical instrument for measuring school performance could be a complex process. It was also acknowledged that this study had its limitations.

7.4 Limitations of This Study

7.4.1 Limitations of Quantitative Research Element of the Study

Although this study provided valuable insights into VAMs in the local context, it is important to

acknowledge the methodological and other limitations. Firstly, this was a small-scale pilot study, and the quantitative sample consisted of only 11 public junior high schools. However, by including class-level data (46 classes), this study was able to reduce the amount of unexplained variance in the models and improve the precision of estimations. Despite this, it was not possible to generalize the findings to a larger population of schools in the city or province, as socioeconomic and other differences may have existed across districts and cities. However, within this local context, the significance of using VAMs to generate more equitable and valid school evaluation results was highlighted, as it enabled the identification of schools that served students with considerable disadvantage in a fairer way.

Secondly, as with any statistical model, there were assumptions and limitations in the data used to calculate value-added score estimates. The quality, reliability, and validity of the data analysed were critical to the value of the results (Thomas & Mortimore, 1996). It was important to note the limitations associated with measurement errors and data imperfections (Goldstein, 1995; OECD, 2008). For example, the student background data was collected through self-reports, which could be subject to inaccuracies due to insincere responses by students (Salim, 2011). However, the author conducted a pilot testing of student questionnaires and provided clear instructions to mitigate this limitation. Additionally, while student attainment data from the LEA was considered a reliable source of information, errors could still exist due to the challenges involved in matching each student's background information with their attainment. The author worked with LEA staff to check for errors, but there was no guarantee of its ultimate quality. Furthermore, in acknowledging the potential benefits of a random slope model for estimating school and class effects, it was important to recognize that the data analysis in this study was limited to employing a random intercept model. Although this limitation was influenced by several factors outlined in Section 2.4.5.2, it was worth noting the necessity to employ a random slope model in future research. Moreover, VAMs only provided a partial picture of school effectiveness and did not account for unmeasured factors outside of school control that may impact student achievement. Therefore, it was important to be aware of caveats and use VAMs in conjunction with other indicators to provide a comprehensive assessment of school performance.

7.4.2 Limitations of Qualitative Research Element of the Study

Similar to the limitations of the quantitative research element, the small number of interviewees in this

study cannot be considered representative of all stakeholders. The study did not explore the perspectives of other stakeholders such as teachers, parents, and students, whose views could have provided a more comprehensive understanding of the implementation of VAMs. However, their perceptions provided a valuable snapshot at one point in time to identify the main factors that may have helped or hindered the implementation of VAMs in the local context. It is worth noting that the LEA played a crucial role in facilitating the selection of headteachers by providing crucial assistance, including introducing the researcher to potential candidates and sharing their contact information. Although they were volunteers, administrative powers may have influenced their decisions to participate to some extent. Therefore, caution should be taken when generalizing the results of the interview data.

The qualitative data collected through interviews may also have been subject to researcher bias. The interpretation and analysis of the data may have been influenced by the researcher's preconceptions, personal beliefs, and values (Merriam, 2009). Therefore, the findings and conclusions derived from the interviews should be interpreted with this in mind. However, to reduce this bias, the researcher conducted the interviews in a neutral and open manner and avoided leading questions. Furthermore, a peer-review process was implemented to enhance the validity and reliability of the qualitative data analysis.

The issue of the researcher's non-native English proficiency should be taken into consideration when interpreting the findings. As a non-native English speaker, the researcher may have encountered difficulties in accurately conveying the intended meanings of participants' responses, which could potentially lead to misunderstandings and misinterpretations of the data. Furthermore, the use of technical terms related to value-added evaluation may not be easily understood by participants who are not familiar with the concepts. However, the researcher took measures to mitigate this limitation, such as conducting pilot testing of the interview questions to ensure clarity and seeking assistance from a colleague to review the interview transcripts in both Chinese and English.

7.5 Future Research

Although this study has made a valuable contribution to our understanding of VAMs in the context of China, as evidenced in Section 7.2, its limitations, as discussed in the preceding section, indicate that

it does not address all relevant questions. Consequently, further research is necessary. To begin with, it is important to conduct studies involving a larger sample of schools outside the scope of this study. This will facilitate an examination of the validity and reliability of VAMs and provide more comprehensive evidence across a wider range of contexts.

Secondly, schools require a comprehensive understanding of their performance across multiple dimensions to effectively plan for school improvement. Therefore, future research should aim to address the limited range of outcomes, specifically academic outcomes, by including a broader range of outcomes such as non-academic subjects and non-cognitive outcomes.

Thirdly, regarding the selection of appropriate value-added models for school evaluation, several studies comparing different value-added models have been conducted to provide policymakers with sufficient evidence (Leckie & Goldstein, 2019; Leckie & Prior, 2022; Marks, 2017; 2021; Thomas, 2001; Thomas & Mortimore, 1996; Timmermans et al., 2011). However, such research is less frequently conducted in the context of China (Yang & Zhang, 2022). Given the varying adjustments made to value-added models, which may include prior attainment, student background characteristics, and class and school context variables, different estimation results may be generated. Therefore, conducting comparative studies of value-added models is valuable to provide evidence for policymakers when selecting appropriate models for implementation.

Furthermore, as demonstrated in Chapter 2, it is essential to incorporate the perspectives of practitioners to enrich our understanding of how VAMs can be effectively implemented. Therefore, conducting a qualitative study involving a broader range of stakeholders, including teachers, inspectorates, and policymakers from city or provincial education authorities, and investigating the factors that must be considered before implementing VAMs in school evaluation practice, may provide a clearer understanding of the effective implementation of VAMs in local contexts. Such research would contribute to enhancing the practical application of VAMs in school evaluation practice.

Finally, it is important to emphasize the role of VAMs as a management tool for improving the quality of school education in the current context of China. Therefore, conducting qualitative case studies would be beneficial in gaining in-depth insights into the reasons for differences in school performance, enhancing our understanding of the advantages of VAMs in improving school performance, and

making a more significant contribution to improving educational practice. In summary, further empirical research is needed in the Chinese context and across different regions to enrich our understanding of VAMs and facilitate their implementation in educational practices, while also adapting to contextual issues in China and beyond.

Reference

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies. *Journal of the Royal Statistical Society*, Series A, 149, 1, pp. 1–43.

Aldridge, A., & Levine, K. (2001). *Surveying the Socia.* World Buckingham: Open.

Australian Curriculum, Assessment and Reporting Authority. (2021). Retrieved from My school. https://www.myschool.edu.au/

Ball, S. (1998). Educational studies, policy entrepreneurship and social theory. In R. S. (Eds.), *School effectiveness for whom? Challenges to the school effectiveness and school improvement movements* (pp. 70-83). London: Falmer Press.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of educational and behavioral statistics*, 29(1), 37-65.

Bee, C. P. (1973). Guidelines for Designing a School Evaluation. *Educational Technology*, 13(5), 44-47.

Bian, Y. F., & Lin, Z.H. (2007). Value-added evaluation: a school evaluation model under the concept of green advancement in education. Journal of Beijing Normal University (Social Science Edition), (06), 11-18.

Blakey, L., & Heath, A. (1992). Differences between comprehensive schools: Some preliminary findings. In D. Reynolds, & P. Cuttance, *School Effectiveness: Research, policy and practice* (pp. 121-133). London: Cassell.

Bollen, K. A., Glanville, J. L., & Stecklov, G. (2001). Socioeconomic status and class in studies of fertility and health in developing countries. *Annual Review of Sociology*, 153-185.

Bondi, L. (1991). Attainment in primary schools. *British Educational Research Journal*, 17,3, pp.203–17.

Bosker, R. J., & Witziers, B. (1996). The magnitude of school effects, or: Does it really matter which school a student attends. In *Annual Meeting of the American Educational Research Association, New York*.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), 77-101.

Brinkmann, S., & Kvale, S. (2005). Confronting the ethics of qualitative research. *Journal of constructivist psychology*, 18(2), 157-181.

Brookover, W. B., Beaty, C., Flood, P., Schweitzer, J., & Wisenbaker, J. (1979). *Schools, Social Systems, and Student Achievement: Schools Can Make a Difference.* New York: Praeger.

Brookover, W. B., Schweitzer, J. H., Schneider, J. M., Beady, C. H., Flood, P. K., & Wisenbaker, J. (1978). Elementary school social climate and school achievement. *American Educational Research Journal*, 15(2), 301–318.

Brown, B., & Saks, D. (1986). Measuring the effects of instructional time on student learning: evidence from the beginning teacher evaluation study. *American Journal of Education*, 94, 480-500.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Bryman, A. (2012). *Social Research Methods (4th.ed).* Oxford, UK: Oxford University Press.

Chapman, C., Reynolds, D., & Muijs, D. (2015). Educational effectiveness and improvement research and practice: The emergence of the discipline. In *The Routledge international handbook of educational effectiveness and improvement* (pp. 33-56). Routledge.

Chen, B., & Dong, J. (2022). The Significance and Strategies of Implementing Value-Added Evaluation under the Background of "Double Reduction". *Education Science Forum*, (32), 76-80.

Chinese Ministry of Education. (2005). *The 9th five-year plan for china's educational development and the development outline by 2010.* Retrived from

Chinese Ministry of Education. (2013). *Opinions on Promoting the Reform of Comprehensive Evaluation of Primary and Secondary Education Quality* . Retrieved from: http://www.moe.gov.cn/srcsite/A06/s3321/201306/t20130608_153185.html

Chinese Ministry of Education. (2017). *Managment standards for schools in basic education.* Retrieved from http://www.moe.gov.cn/srcsite/A06/s3321/201712/t20171211_321026.html

Coe, R., & Fitz-Gibbon, C. T. (1998). School effectiveness research: Criticisms and recommendations. *Oxford Review of Education*, 24(4), 421 – 438.

Coleman, J. S., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity.* Government Printing Office: Washington, DC.

Crawford, C., Dearden, L., & Meghir, C. (2007). *When you are born matters: The impact of date of birth on child cognitive outcomes in England.* (IFS External Reports). Institute for Fiscal Studies:London, UK.

Creemers, B. (1992). 'School effectiveness and effective instruction—The need for a further relationship'. *School effectiveness and improvement*, 37-48.

Creemers, B. P. (1994). *The Effective Classroom.* London: Cassell.

Creemers, B. P. (2005). Educational effectiveness the development of the field. *Keynote address presented at the first International Conference on School Effectiveness and School Improvement.*

Creemers, B. P. (2006). The importance and perspectives of international studies in educational effectiveness. *Educational Research and Evaluation*, 12(6): 499–511.

Creemers, B. P. M., and Kyriakides L. (2008) *The Dynamics of Educational Effectiveness: A Contribution to Policy Practice and Theory in Contemporary Schools.* London: Routledge.

Creemers, B. P., Kyriakides, L., & Sammons, P. (2010). *Methodological Advances in Educational Effectiveness* . London/New York: Routledge.

Creemers, B., & Kyriakides, L. (2016). Theory development in educational effectiveness research. In *The Routledge international handbook of educational effectiveness and improvement: Research, policy, and practice* (pp. 149-172). Routledge.

Creemers, B., & Reynolds, D. (1990). School effectiveness and school improvement: A mission statement. *School Effectiveness and School Improvement*, 1,1, pp.1-3.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches (2nd ed.).* Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2009). Mapping the field of mixed methods research. *Journal of mixed methods research*, 3(2), 95-108.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches.* Sage publications.

Creswell, J., & Clark, V. L. (2011). *Designing And Conducting Mixed Methods Research (2nd Edition).* USA: California: Sage.

De Fraine, B., Van Damme, J., Van Landeghem, G., Opdenakker, M. C., & Onghena, P. (2003). The effect of schools and classes on language achievement. *British educational research journal*, 29(6), 841-859.

De Jong, R., Westerhof, K., & Kruiter, J. (2004). Empirical evidence of a compre - hensive model of school effectiveness: A multilevel study in mathematics in the 1st year of junior general education in the Netherlands. *School Effectiveness and School Improvement*, 15(1): 3–31.

Dellinger, A. B., & Leech, N. L. (2007). Toward a unified validation framework in mixed methods research. *Journal of mixed methods research*, 1(4), 309-332.

Department for Education. (2010). *The importance of teaching: The schools white paper 2010.* Retrieved from https://www.gov.uk/government/publications/the-importance-of-teaching-the-schools-white-paper-2010

Department for Education. (2020). *Secondary accountability measures: Guide for maintained secondary schools; academies, and free school.* Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/872997/Secondary_accountability_measures_guidance_February_2020_3.pdf

Dong, J. (2021). Using value-added evaluation to promote educational equity and quality improvement-- A case study of value-added evaluation on academic performance of senior high schools in Baotou City. *Inner Mongolia Education* , (13),28-34.

Du, P., & Yang, Z. (2011). Value-added Evaluation of School Effectiveness in Rural Junior Middle Schools- Empirical Study Based on Data from Five Provinces in West of China. *Journal of Beijing Normal University (Social Sciences)* , (06), 91-97.

Du, X., & Hao, C. (2021). A Study on Value-Added Evaluation of Primary and Secondary School Students. *Education and Teaching Forum*, (29), 13-16.

Easen, P., & Bolden, D. (2005). 'Location, Location, Location: What Do League Tables Really Tell Us about. *Education*, 3-13.

Edmonds, R. R. (1979). Effective schools for the urban poor. *Educational Leadership*, 37, 15 27.

Elliot, K., Smees, R., & Thomas, S. (1998). aking the most of your data: school self-evaluation using value added measures. *Improving Schools*, 1(3), 59-67.

Fan, M., & Gao, L. (2019). Value-added Evaluation of Senior High School Education and Teaching Efficiency Based on the Score Data of Zhongkao and Gaokao. *Journal of China Examinations*, (10),6-13.

Feng, B., & Zhou, Y. (2022). A Review of Some Controversial Issue on Value-added Evaluation in China. *Journal of Shanghai Educational Research*, (01),66-72.

Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International journal of qualitative methods*, 5(1), 80-92.

Fertig, M. (2000). Old wine in new bottles? Researching effective schools in developing countries. *School Effectiveness and School Improvement*, 11(3), 385-403.

Fitz-Gibbon, C. (1991). 'Multi-level modeling in an indicator system'. In S. a. Raudenbush, *Schools, Classrooms, and Pupils: International Studies of Schooling from a Multilevel Perspective* (pp. 67-84). San Diego: Academic Press.

Fitz-Gibbon, C. (1995). *The Value Added Ntional Project General Report: Issues to be Considered in the Design of a National Value-Added System.* London: The School Curriculum and Assessment Authority.

Foley, B., & Goldstein, H. (2012). *Measuring success: League tables in the public sector.* The British Academy.

Gao, C., Huang, H., & li, L. (2021). Exploring Value-added Evaluation in Regions and Improving School Effectiveness. *China Modern Educational Equipment*, (20),11-12.

Gao, R., Wang, B., & Lin, W. (2006). The foundation of moral theory- Research and reflection on leadership theory in China. *Science and Technology Management Research*, (6), 142-147.

Gao, X., & Song, N. (2021). Analysis on the Promotion of Value-Added Evaluation to the High-Quality Development of Basic Education in China. *Journal of Jiangxi Normal University (Philosophy and Social Sciences Edition)* , (06), 100-106.

Golden-Bibble, K., & Locke, K. (1993). Appealing works: an investigation of how ethnographic texts convince. *Organization Science*, 4(4), 595-616.

Goldstein, H. (1991). 'Better ways to compare schools?'. *Journal of Educational Statistics*, 16,2, pp.89–92.

Goldstein, H. (1993). *Interpreting International Comparisons of Student Achievement.* Paris:UNESCO.

Goldstein, H. (1995). Hierarchical data modeling in the social sciences. *Journal of Educational and Behavioral Statistics*, 20(2): 201–4.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, 8(4), 369–395.

Goldstein, H. (2003). Multilevel modelling of educational data. . *Methodology and epistemology of multilevel analysis*, 2, 25-41.

Goldstein, H. (2011). *Multilevel statistical models (4th edn)* . John Wiley & Sons.

Goldstein, H., & Leckie, G. (2008). School league tables: What can they really tell us? *Significance*, 5(2): 67–9.

Goldstein, H., & Myers, K. (1997). School effectiveness research: a bandwagon, a hijack or a journey towards enlightenment? *British Educational Research Association Conference.* University of York.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. J*ournal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(3), 385-409.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding statistics: statistical issues in psychology, education, and the social sciences*, *1*(4), 223-231.

Goldstein, H., Leckie, G., & Prior, L. (2020). Providing educational accountability for Local Authorities based up sampling pupils within schools: moving away from simplistic school league tables. *arXiv preprint: 2002.09897*.

Grace, G. (1998). Realizing the mission: Catholic approaches to school effectiveness. . In R. S. (Eds.), *School effectiveness to whom? Challenges to the school effectiveness and school improvement movements* (pp. 117-127). London: Falmer Press.

Gray, J., Jesson, D., & Simes, N. (1990). Estimating Differences in the Examination Performance of secondary schools in six LEAs- A multilevel approach to school effectiveness. *Oxford Review of Education*, 16 (2) 137-156.

Guan, Z. (2022). The Influence of Private Tutoring on Academic Performance of Junior High School Students: Based on PSM-DID Method Estimation. *China Economics of Education Review*, 7(1),58-74.

Guo, Y., & Wang, Q. (2021). Knowledge Graph and Prospect of Value-added Evaluation Research. *Educational Measurement and Evaluation*, (07),3-10.

Hadfield, M., & Chapman, C. (2015). Qualitative methods in educational effectiveness and improvement research. In *The Routledge International Handbook of Educational Effectiveness and Improvement* (pp. 202-219). Routledge.

Hall, J., Lindorff, A., & Sammons, P. (2020). *International perspectives in educational effectiveness research.* Cham: Springer.

Harber, C., & Muthukrishna, N. (2000). School effectiveness and school improvement in context: The case of South Africa. *School effectiveness and school improvement*, 11(4), 421-434.

Hauser, R. (1970). Context and consex: A cautionary tale. *American Journal of Sociology*, 75, pp.645–54.

Hauser, R. (1971). *Socioeconomic Background and Educational Performance.* Washington,DC: Arnold M. Rose Series, American Sociological Association.

Haworth, C., Asbury, K., Dale, P., & Plomin, R. (2011). Added value measures in education show genetic as well as environmental influence. *PLoS ONE* , 6(2):E16006.

Hennink, M., Hutter, I., & Bailey, A. (2020). *Qualitative research methods.* London: Sage Publications.

Heyneman, S., & Loxley, W. (1983). The effect of primary school quality on academic achievement across twenty-nine high and low income countries. *American Journal of Sociology*, 88, 6, pp.1162–94.

Hibpshman, T. (2004). A review of value-added models. *Kentucky Education Professional Standards Board*.

Hill, P., Rowe, K., & Holmes-Smith, P. (1995). *Factors Affecting Students' Educational Progress: Multilevel Modelling of Educational Effectiveness.* Leeuwarden, the Netherlands: Paper presented at the 8th International Congress for School Effectiveness and Improvement.

Howitt, D. (2019). *Introduction to qualitative research methods in psychology.* Pearson UK.

Hu, Z. Q., Zhong, Y., & Wang, J. (2022). Construction and Practice of Value-Added Evaluation Model for Academic Achievement of Primary and Secondary School Students - Based on the Research of Guangzhou City's Sunshine Evaluation in Compulsory Education. *Education Theory and Practice* , (11), 18-22.

Ihantola, E. M., & Kihn, L. A. (2011). Threats to validity and reliability in mixed methods accounting research. *Qualitative Research in Accounting & Management*, 8(1), 39-58.

Jencks, C. S., Smith, M., Ackland, H., Bane, M. J., Cohen, D., Gintis, H., . . . Michelson, S. (1972). *Inequality, a reassessment of the effects of family and schooling in America.* New York:Basic.

Jesson, D. (1996). Value Added Measures of School GCSE Performance: An Investigation Into the Role of Key Stage 3 Assessments in Schools: an Interim Report. Department for Education and Employment.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. *Educational Researcher*, 14-26.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. . *Journal of mixed methods research*, 1(2), 112-133.

Keeves, J. P., Hungi, N., & Afrassa, T. (2005). Measuring value added effects across schools: Should schools be compared in performance? *Studies in Educational Evaluation*, 31(2–3), 247–266.

Kelly, A. (2012). Measuring equity and equitability in school effectiveness research. *British Educational Research Journal*, 38(6): 977–1002.

Koretz, D. (2017). *The testing charade: Pretending to make schools better.* University of Chicago Press.

Kyriakides, L. (2005). Extending the comprehensive model of educational effectiveness by an empirical investigation. *School Effectiveness and School Improvement*, 16(2): 103–52.

Kyriakides, L., & Charalambous, C. (2004). Extending the scope of analysing data of IEA studies: Applying multilevel modelling techniques to analyse TIMSS data. Paper presented at the First IEA International Research Conference (IRC-2004), Nicosia,Cyprus, 11–13 May.

Kyriakides, L., & Creemers, B. P. (2008). A longitudinal study on the stability over time of school and teacher effects on student outcomes. *Oxford review of education*, 34(5), 521-545.

Kyriakides, L., Campbell, R., & Gagatsis, A. (2000). The significance of the classroom effect in primary schools: An application of Creemers' comprehensive model of educational effectiveness. *School Effectiveness and School Improvement*, 11(4): 501–29.

Leckie, G., & Browne, B. (2021). Introduction to Multilevel Modelling Using MLwiN, R, or Stata MLwiN Practicals. Centre for Multilevel Modelling, University of Bristol: Centre for Multilevel Modelling.

Leckie, G., & Goldstein, H. (2017). The evolution of school league tables in England 1992–2016:'Contextual value-added','expected progress' and 'progress 8'. *British Educational Research Journal, 43(2)*, 193-212.

Leckie, G., & Goldstein, H. (2019). The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England. *British Educational Research Journal*, 45(3), 518-537.

Leckie, G., & Prior, L. (2022). A comparison of value-added models for school accountability. *Schol Effectiveness and School Improvement*, 33:3, 431-455.

Lenkeit, J. (2013). Effectiveness measures for cross-sectional studies: A comparison of value-added models and contextualised attainment models. *School Effectiveness and School Improvement*, 24 (1), 1–25.

Levine, D., & Lezotte, L. (1990). *Unusually Effective Schools: A Review and Analysis of Research and Practice.* Madison, WI: National Center for Effective Schools Research and Development.

Li, Y., Zhen, P., & Huang, Y. (2022). Value-Added Evaluation in Basic Education: Connotation, Value, and Empirical Research - Taking the Reading Literacy of Primary and Secondary School Students in the New Era as an Example. *China Educational Technology*, (10), 47-55+71.

Liang, X., Kidwai, H., & Zhang, M. (2016). *How Shanghai does it: Insights and lessons from the highest-ranking education system in the world.* Retrieved from World Bank: https://openknowledge.worldbank.org/bitstream/handle/10986/24000/9781464807909.pdf

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry.* Sage.

Lindorff, A., Sammons, P., & Hall, J. (2020). International perspectives in educational effectiveness research: A historical overview. In *International perspectives in educational effectiveness research* (pp. 9-31). Springer, Cham.

Liu, Y. M. (2022). An Empirical Study of Junior High School Mathematics Value-Added Assessment Based on Regional Data. *Journal of Shanghai Educational Research*,(08), 34-39

Liu, S., & Teddlie, C. (2003). The ongoing development of teacher evaluation and curriculum reform in the People's Republic of China. *Journal of Personnel Evaluation in Education*, 17(3), 243-261.

Liu, X. Z., & Tian, X. (2020). The Direction Change of Students'Evaluation Reform of Basic Education in the New Era. *China Examinations*, (8)16-19.

Luyten, H., & Sammons, P. (2010). Multilevel modelling. In *Methodological advances in educational effectiveness research* (pp. 260-290). Routledge.

Luyten, H., Visscher, A., & Witziers, B. (2005). School effectiveness research: From a review of the criticism to recommendations for further development. *School effectiveness and school improvement*, 16(3), 249-279.

Lv, W. B. (2015). *An Empirical Study on the Effectiveness of Junior High School Based on Value-added Evaluation.* Master Dissertation, Shangdong Normal University.

Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. *A Journal of the Humanities & Social Sciences*, 13(1).

Ma, X. Q. (2020). Exploring Value-added Evaluation, What are Our Concerns? *Journal of Primary and Middle School Management*, (10),5-7.

Ma, X. Q. (2021). Design and Research on the Evaluation System of Primary and Secondary School Students' Growth Record Bags under the Background of the New Curriculum. *New Curriculum*, (48),1.

Ma, X., & Peng, W. (2006). Value-Added Evaluation of School Effectiveness. *Journal of Education*, (10).

Marks, G. N. (2017). Is adjusting for prior achievement sufficient for school effectiveness studies? *Educational Research and Evaluation*, 23(5–6), 148–162.

Marks, G. N. (2021). Should value-added school effects models include student- and school-level covariates? Evidence from Australian population assessment data. *British Educational Research*, 47(1), 181–204.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica, CA: The RAND Corporation.

McPherson, A. (1992). 'Measuring value added in schools'. *National Commission on Education Briefing No. 1.* , London: NCE.

Merriam, S. B. (2009). *Qualitative Research: A Guide to Design and Implementation.* Jossey-Bass.

Ministry of Human Resources and Social Security of the People's Republic of China. (2015). *Occupational Classification Code of the People's Republic of China.* China Labor and Social Security Press.

Mohajan, H. K. (2017). Two criteria for good measurements in research: Validity and reliability. *Annals of Spiru Haret University. Economic Series*, 17(4), 59-82.

Mokonzi, G. B., Van Damme, J., De Fraine, B., Vitamara, P. M., Kimbuani, G. M., Mukiekie, A., & ... Bela, J. P. (2020). Educational effectiveness research in Africa: The case of the Democratic Republic of the Congo (DRC). In *International Perspectives in Educational Effectiveness Research* (pp. 185-207). Springer, Cham.

Monk, D. (1992). Education productivity research: An update and assessment of its role in education finance reform . *Educational Evaluation and Policy Analysis*, 14(4): 307–32.

Mortimore, P. (1991). School effectiveness research: Which way at the crossroads? . *School Effectiveness and School Improvement*, 2(3): 213–29.

Mortimore, P. (1998). *The road to improvement: Reflections on school effectiveness.* Abington, UK: Taylor & Francis.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., & Ecob, R. (1988). *School Matters.* London, UK: Open Books.

Muijs, D., & Brookman, A. (2015). Quantitative methods. In *The Routledge international handbook of educational effectiveness and improvement* (pp. 205-233). Routledge.

Muijs, D., & Reynolds, D. (2000). School effectiveness and teacher effectiveness in mathematics: Some preliminary findings from the evaluation of the mathematics enhancement programme (primary). *School Effectiveness and School Improvement*, 11(3): 273–303.

Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H., & Earl, L. (2014). State of the art: Teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, 25(2): 231–56.

Muñoz-Chereau, B. (2013). *Searching for fairer ways of comparing Chilean secondary schools performance: a mixed methods study investigating contextual value added approaches.* PhD Thesis. University of Bristol.

Muñoz-Chereau, B., & Thomas, S. M. (2016). Educational effectiveness in Chilean secondary education: comparing different 'value added'approaches to evaluate schools. *Assessment in Education: Principles, Policy & Practice*, 23(1), 26-52.

Munoz-Chereau, B., Anwandter, A., & Thomas, S. (2020). Value-added indicators for a fairer Chilean school accountability system: A pending subject. *Journal of Education Policy*, 35(5), 602–622.

Murnane, R. (1981). 'Interpreting the evidence on school effectiveness'. *Teachers College Record*, 83,19-35.

National Bureau of Statistics. (2021). *Major Figures on 2020 Population Census of China.* Retrieved from http://www.stats.gov.cn/sj/pcsj/rkpc/d7c/202303/P020230301403217959330.pdf

Nuttall, D. (1991). An instrument to be honed. *Times Educational Supplement*, 13.

OECD. (2007). *PISA 2006: Science Competencies for Tomorrow's World.* Paris: OECD

OECD. (2008). *Measuring Improvements in Learning Outcomes Best Practices to Assess the Value-Added of Schools.* Organisation for Economic Co-operation and Development Publishing & Centre for Educational.

OECD. (2009). *School Evaluation: Current Practices in OECD Countries and a Literature Review.* Organisation for Economic Co-operation and Development..

OECD. (2013). *Synergies for Better Learning: an internaitonal perspective on evaluation and assessment.* Organisation for Economic Co-operation and Development Publishing & Centre for Educational.

OECD. (2019). *PISA 2018: Assessment and Analytical Framework.* OECD Publishing Paris.

OECD. (2020). *Benchmarking the Performance of China's Education System, PISA.* OECD Publishing, Paris.

Onwuegbuzie, A. J., & Johnson, R. B. (2006). The validity issue in mixed research. *Research in the Schools*, 13(1), 48-63.

Opdenakker, M.-C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, 11(1): 103–30.

Peng, P., Hochweber, J., & Klieme, E. (2013). Test Score or Student Progress? A Value-Added Evaluation of School Effectiveness in Urban China. *Frontiers of Education in China*, 8(3), 360-377.

Peng, S., & Zhang, H. (2021). Theoretical Thinking and Practical Exploration of Educational Value-Added Evaluation. *Science Education and Culture (Mid-monthly)*, (12), 10-12.

Peng, W., Thomas, S., Yang, X., & Li, J. (2006). Developing school evaluation methods to improve the quality of schooling in China: a pilot 'value added' study. *Assessment in Education: Principles, Policy & Practice*, 13:2, 135-154.

Plowden, B. (1967). *Children and their Primary Schools: A Report of the Central Advisory Council for Education (England).* London: HMSO.

Prior, L., Jerrim, J., Thomson, D., & Leckie, G. (2021). A review and evaluation of secondary school accountability in England: Statistical strengths, weaknesses and challenges for 'Progress 8'raised by COVID-19. *Review of Education*, *9*(3), e3299.

Raudenbush, S., & Bryk, A. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59, pp.1–17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.).* Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Willms, J. D. (1996). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 10,4,pp.307-35.

Ray, A. (2006). School value added measures in England. *A paper for the OECD Project on the Development of Value-Added Models in Education Systems*.

Rea, J., & Weiner, G. (1998). Cultures of blame and redemption—When empowerment becomes control: Practitioners' views of the effective school movement. In R. S. (Eds.), *School effectiveness for whom? Challenges to the school effectiveness and school improvement movements* (pp. 21-32). London: Falmer Press.

Reezigt, G. J., Guldemond, H., & Creemers, B. P. (1999). Empirical validity for a comprehensive model on educational effectiveness. *School Effectiveness and School Improvement*, 10(2): 193–216.

Ren, Y. (2022). Effective or Biased: The Questions Raised by Teacher Value-Added Evaluation in the United States and the Implications for Exploring Value-Added Evaluation in China. *China Examination*, (04), 34-43.

Reynolds, D. (1988). British school improvement research: The contribution of qualitative studies. *International Journal of Qualitative Studies in Education*, 1, 2, pp. 143–54.

Reynolds, D. (2002). School effectiveness: The international dimension. In *The international handbook of school effectiveness research* (pp. 246-270). Routledge.

Reynolds, D., & Stoll, L. (1996). Merging the school effectiveness and school improvement: the knowledge bases. In R. Bollen, B. P. Creemers, D. Hopkins, N. Lagerweij, D. Reynolds, & L. Stoll, *Making Good Schools: Linking School Effectiveness and Improvement* (pp. 95-153). Routledge.

Reynolds, D., & Teddlie, C. (2002). An introduction to school effectiveness research. In *The international handbook of school effectiveness research* (pp. 17-39). Routledge.

Reynolds, D., Kelly, A., Harris, A., Jones, M., Adams, D., Miao, Z., & Bokhove, C. (2020). Extending educational effectiveness: A critical review of research approaches in international

effectiveness research, and proposals to improve them. In *International Perspectives in Educational Effectiveness Research* (pp. 121-145). Springer.

Reynolds, D., Sammons, P., de Fraine, B., van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, 25(2): 197–230.

Reynolds, D., Sammons, P., Stoll, L., Barber, M., & Hillman, J. (1996). School effectiveness and school improvement in the United Kingdom. *School effectiveness and school improvement*, 7(2), 133-158.

Rubin, A., & Babbie, E. (2017). *Research Methods for Social Work.* Boston: MA: Cengage Learning.

Rutter, M., Maughan, B., Mortimore, P., Ouston, J., & Smith, A. (1979). *Fifteen Thousand Hours: Secondary Schools and Their Effects on Children.* London: Open Books and Boston, MA: Harvard University Press.

Salim, M. M. (2011). *Exploring issues of school effectiveness and self-evaluation at the system and school levels in the context of Zanzibar.* Doctoral dissertation, University of Bristol.

Sammons, P. (1999). School effectiveness: Coming of age in the 21st century. *Management in Education*, 13(5), 10-13.

Sammons, P. (2010). The contribution of mixed methods to recent research on educational effectiveness. In *handbook of mixed methods in social and behavioral research* (pp. 697-723). Sage.

Sammons, P., Davis, S., & Gray, J. (2016). Methodological and scientific properties of school effectiveness research: Exploring the underpinnings, evolution, and future directions of the field. In *The Routledge international handbook of educational effectiveness and improvement* (pp. 25-76). Routledge.

Sammons, P., Hillman, J., & Mortimore, P. (1995). *Key characteristics of effective schools: A review of school effectiveness research.* London: OFSTED.

Sammons, P., Mortimore, P., & Thomas, S. M. (1996). Do schools perform consistently across outcomes and areas. In *Merging traditions: The future of research on school effectiveness and school improvement* (pp. 3-29). Cassell.

Sammons, P., Thomas, S., Mortimore, P., Walker, A., Cairns, R., & Bausor, J. (1998). Understanding differences in academic effectiveness: practitioners' views. *School Effectiveness and School Improvement*, 9(3), 286-309.

Saunders, L. (1999). *Value added measurement of school effectiveness: A critical review.* Slough: National Foundation for Educational Research.

Saunders, L. (2000). Understanding schools' use of 'value added'data: the psychology and sociology of numbers. *Research papers in education*, 15(3), 241-258.

Saunders, L. (2001). A science in the service of an art? The use of 'value added'analyses of school performance to aid school improvement.

Scheerens, J. (1990). School Effectiveness and the Development of Process Indicators of School Functioning. *School Effectiveness and School Improvement*, 61-80.

Scheerens, J. (1992). *Effective Schooling, Research, Theory and Practice.* London: Cassell.

Scheerens, J. (1997). Conceptual models in theory embedded principles on effective schooling. *School Effectiveness and School Improvement*, 8, 3, pp.269–310.

Scheerens, J. (2005). Review of school and instructional effectiveness research. *Paper commissioned for the EFA Global Monitoring Report.*

Scheerens, J. (2013). What Is Effective Schooling? A review of current though and practice. *Report prepared for International Baccalaureate Organization.*

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectivenes.* Oxford/New York/Tokyo: Pergamon.

Scheerens, J., & Creemers, B. (1989). Conceptualizing school effectiveness. In: B.P.M. Creemers & J. Scheerens (Eds.), Developments in school effectiveness research. *Special issue of the International Journal of Educational Research*, (13)7.

Scheerens, J., Glas, C. A., & Thomas, S. (2003). *Educational evaluation, assessment, and monitoring: A systemic approach* (Vol. 13). Taylor & Francis.

Shao, Y., Shao, Y.Y., & Liu, J. (2021). Value-added Assessment Focuses on Schools that Bring More Benefits to Students-Taking Empirical Data of One District in Beijing as an Example. *China Examinations*, (09),40-45.

Sirin, S. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research*, 75(3), 417-453.

State Council. (2015). *National plan for monitoring compulsory education quality.*

State Council. (2020). Comprehensive Plan for Deepening the Educational Assessment Reform in the New Era. Retrieved from http://www.gov.cn/gongbao/content/2020/content_5554488.htm

State Council. (2021) Opinions on Further Reducing the Homework and Extracurricular Training Burden on Students in the Compulsory Education Stage. Retrieved from http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202107/t20210724_546576.html

Stoll, L. (1996). Linking school effectiveness and school improvement: Issues and possibilities. In J. Gray, D. Reynolds, C. Fitz-Gibbon, & D. Jesson, *Merging Traditions: The Future of Research on School Effectiveness and School Improvement* (pp. 51-73). London: Cassell.

Stoll, L., & Fink, D. (1992). Effecting school change: The Halton approach. *School Effectiveness and School Improvement*, 3, 1, pp.19–41.

Stoll, L., & Mortimore, P. (1995). *School Efffectiveness and School Improvement.* Viewpoint Number 2,London：University of London Institute of Education.

Stoll, L., Mcmahon, A., & Thomas, S. (2006). Identifying and leading effective professional learning communities. *Journal of School Leadership*, 16(5), 611-623.

Strand, S. (1998). A Value added'analysis of the 1996 primary school performance tables. *Educational Research*, 40(2), 123-137.

Tang, L. (2005). *A study on school effectiveness evaluation.* Dissertation presented to the faculty in the Department of Educational Administration, East China Normal University, Shanghai, China.

Tashakkori, A., & Teddlie, C. (1998). *Mixed Methods: Combining the Qualitative and Quantitative Approaches.* Thousand Oaks, CA: Sage.

Teddlie, C. (1994). 'Integrating classroom and school data in school effectiveness research'. In D. e. Reynolds, *Advances in School Effectiveness Research and Practice* (pp. 111-132). Oxford:Pergamon.

Teddlie, C. B., & Tashakkori, A. (2009). *Foundations of Mixed Methods Research: Integrating Quantitative and Qualitative Approaches to Social and Behavioural Sciences.* Los Angeles, CA: Sage.

Teddlie, C., & Liu, S. (2008). Examining teacher effectiveness within differentially effective primary schools in the People's Republic of China. *School Effectiveness and School Improvement*, 19(4), 387-407.

Teddlie, C., & Reynolds, D. (2002). Current topics and approaches in school effectiveness research: The contemporary field. In *The international handbook of school effectiveness research* (pp. 40-66). Routledge.

Teddlie, C., & Reynolds, D. (2002). *International handbook of school effectiveness research.* London, UK: Falmer.

Teddlie, C., & Sammons, P. (2010). Applications of mixed methods to the field of educational effectiveness research. In *Methodological advances in educational effectiveness research* (pp. 129-166). Routledge.

Teddlie, C., & Stringfield, S. (1993). *Schools Make a Difference: Lessons Learned from a 10-year Study of School Effects.* New York: Teachers College Press.

Teddlie, C., Stringfield, S., & Reynolds, D. (2002). Context issues within school effectiveness research. In *In The international handbook of school effectiveness research* (pp. 174-200). Routledge.

Thomas, S. (1998). Value-added measures of school effectiveness in the United Kingdom. *Prospects*, 28(1), 91-108.

Thomas, S. (2001). Dimensions of secondary school effectiveness: Comparative analyses across regions. *School Effectiveness and School Improvement*, 12(3), 285–322.

Thomas, S. (2005). Adoption of "value-added" Evaluation Indicators for Assessing School Performance. *Educational Research*, (09),20-27.

Thomas, S. M. (2020). School and teacher value added performance and the relationship with teacher professional development in mainland China. In *International Perspectives in Educational Effectiveness Research* (pp. 209-29). Springer, Cham.

Thomas, S., & Mortimore, P. (1996). Comparison of value-added models for secondary school effectiveness. *Research papers in education*, 11(1), 5-33.

Thomas, S., & Nuttall, D. (1993). *An Analysis of 1992 Key Stage 1 Results in Lancashire—Final Report: A Multilevel Analysis of Total Subject Score, English Score and Mathematics Score.* London: Institute of Eduzation.

Thomas, S., & Peng, W.-J. (2011). Methods to Evaluate Educational Quality and Improvement in China. Chapter 10 . In J. R. (Ed), *Understanding China's Education reform: Creating cross cultural knowledge, pedagogies and dialogue* (pp. 75-91). Routledge.

Thomas, S., Peng, W. J., & Gray, J. (2007). Modelling patterns of improvement over time: Value added trends in English secondary school performance across ten cohorts. *Oxford Review of Education*, 33(3): 261–295.

Thomas, S., Peng, W., Tian, H., Li, J., Ren, C., & Ma, X. (2012). Research on Value Added Evaluation of School Effectiveness. *Educational Research*, (07),29-35.

Thomas, S., Peng, W.-J., & Li, J. (2015). Time Trends in School Value Added Performance and the Relationship between Value Added and Teachers Professional Development in China. *Educational Research*, 35, 7, p. 64-72 .

Thomas, S., Sammons, P., & Mortimore, P. (1994). *Stability in secondary schools: Effects on student GCSE outcomes.* Oxford: Paper presented at the Annual Conference of the British Educational Research Association.

Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997). Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years. *School effectiveness and school improvement*, pp. 8(2), 169-197.

Thrupp, M. (1999). *Schools making a difference: school mix, school effectiveness, and the social limits of reform.* McGraw-Hill Education (UK).

Thrupp, M. (2001). Sociological and political concerns about school effectiveness research: Time for a new research agenda. *School effectiveness and school improvement*, 12(1), 7-40.

Thrupp, M. (2007). Education's 'inconvenient truth', Part one: Persistent middle class advantage. *New Zealand Journal of Teacher's Work*, 4(2): 77–88.

Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research*, 37, 483–504.

Timmermans, A. C., & Thomas, S. M. (2015). The impact of student composition on schools' value added performance: A comparison of seven empirical studies. *School Effectiveness and School Improvement*, 26(3), 487–498.

Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement*, 22(4), 393–413.

Townsend, T., MacBeath, J., & Bogotch, I. (2015). Critical and alternative perspectives on educational effectiveness and improvement research. In *The Routledge International Handbook of Educational Effectiveness and Improvement* (pp. 412-439). Routledge.

Tranmer, M., & Steel, G.D. (2001). Ignoring a level in a multilevel model: evidence from UK census data. Environment and Planning A, 33(5), 941-948.

Trower, P., & Vincent, L. (1995). *The Value Added National Project Technical report: Secondary.* London: School Curriculum and Assessment Authority.

Tucker, M. (2011). *Standing on the shoulders of giants: An American agenda for education reform.* Retrieved from http://www.ncee.org/wp-content/uploads/2011/05/Standing-on-the-Shoulders-of-Giants-An-American-Agenda-for-Education-Reform

Tymms, P. (1997). *The Value-Added National Project. Technical Report: Primary 4. Value-Added Key Stage 1 to Key Stage 2.* SCAA Publications.

Tymms, P., Merrell, C., Heron, T., Jones, P. A., & Henderson, B. (2008). The importance of districts. *School Effectiveness and School Improvement*, 19(3): 261–74.

UNESCO. (2005). *Education for all global monitoring report: The quality imperative.*

van de Grift, W. (1990). Educational leadership and academic achievement in secondary education. *School Effectiveness and School Improvement*, 1, 1, pp.26–40.

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. *BMC medical research methodology*, 18, 1-18.

Wang, G. Q. (2018). Value-Added Evaluation Based on Regional Four-Dimensional Evaluation System in Junior High Schools. *Educational Measurement and Evaluation*, (03),34-38.

Wang, J., Dai, H., & Zhou, Y. (2009). An empirical study of educational value-added assessment-Taking 30 Senior High Schools in Shangrao City Jiangxi Province as An Example. *China Examinations*, (009), 3-9.

Wang, T. P., & Pai, D. Q. (2022). Education Value-Added Evaluation Reform in the New Era: From Data Description to Value Presentation. *Journal of China Examinations*, (10) 13-18.

Weber, G. (1971). *Inner city children can be taught to read: Four successful schools.* Washington, DC: Council for Basic Education.

Wen, S., & Sun, G. (2022). Value-Added Evaluation: Facilitating Sustainable Development for Every Student. *Journal of Shanghai Educational Research*, (3),70-75.

Willms, J. D. (1992). *Monitoring School Performance: A Guide for Educators.* London: Falmer Press.

Wynne, E. A. (1981). Looking at good schools. *The Phi Delta Kappan*, 62(5), 377-381.

Xie, X., & Zhang, H. (2021). The Development Dilemma and Solution Strategy of Students' Value-added Assessment from the Perspective of Five-dimension Educating Integration. *China Educational Technology*, (11),32-38.

Xin, J. (2016). Value-added evaluation: Promoting the sustainable development of every school. *Educational Measurement and Evaluation*, (11), 1.

Xin, T. (2019). Deepening the reform of educational evaluation and establishing a sound evaluation system. *Tsinghua University Journal of Education*, (01), 8-10.

Xin, T. (2020). Exploring Value-added Evaluation, Several Key Questions. *Journal of Primary and Middle School Management*, (10),1.

Xin, T., & Li, G. (2020). Functional Positioning and Key Contents of Educational Quality Evaluation in the Era of High-Quality Development. *People's Education*, (20), 16-18.

Xin, T., Jiang, Y., & Liu, W. L. (2012). Value-Added Evaluation of Senior High School- An Empirical Study Based on the Score Data of Zhongkao and Gaokao from 40 Senior High Schools. *Administration of Elementary and Secondary Schools*, (6), 4-7.

Yang, M., Goldstein, H., Rath, T., & Hill, N. (1999). The use of assessment data for school improvement purposes. *Oxford Review of education*, 25(4), 469-483.

Yang, Z., & Zhang, F. (2022). Hierarchical Linear Modeling and Its Application in Value-Added Evaluation. *Educational Measurement and Evaluation*, (02), 3-11.

Yu, G., & Thomas, S. M. (2008). Exploring school effects across southern and eastern African school systems and in Tanzania. *Assessment in Education: Principles, Policy & Practice*, 15(3), 283-305.

Zhan, Q. (2001). On school effectiveness research. *Theory and Practice of Education*, 21(6), 25–28.

Zhang, L. (2016). *School and teacher effectiveness of senior high schools in western China.* (Doctoral dissertation, University of Bristol).

Zhang, X. (1998). Introduction to value added for quality education improvement. *Journal of Educational Development*, 2/3, 14, 81.

Zhang, Z., Wang, L., & Ji, K. (2022). Empowering the Transformation of Education Evaluation in the New Era with Big Data: Technological Logic, Realistic Dilemmas, and Implementation Paths. *Educational Technology Research*, (05), 33-39.

Zhen, H. (2019). *The role of the school inspection system in demonstrating and improving the quality of compulsory education: exploring stakeholder perceptions in Shandong Province China.* Doctoral dissertation, University of Bristol.

Zhen, Z., & Song, N. (2021). The Rationales of Value-added evaluation of Basic Education in New Era. *Research in Educational Development*, (10),1-7-17.

Appendix 1 Request Permission Letter

Director of Education Bureau of W District
Education Bureau of W District
China

Subject: Request for Permission and Support to Undertake Research on" Exploring the feasibility of value-added measures as an alternative method to measure public junior high school performance in the context of China."

Please refer to the subject mentioned above.

I am currently an EdD student at the Graduate School of Education. University of Bristol, UK. As the dissertation's request, I am required to undertake a piece of research that will finally be presented in the form of a dissertation.

This letter is to request the Education Bureau of W District to grant me permission to conduct the research mentioned above. This study aims to explore the feasibility of the value-added approach as an alternative method to improve school effectiveness evaluation in the local context of China. The research will carry out a questionnaire survey that will involve all Grade 9 students in public junior high schools in the W district in 2021. This questionnaire will collect the students' basic information about their family as well as themselves. Collecting this data will help the researcher carry out the multilevel analysis to estimate school effects on students' High School Entrance Exam performance. The research also involves interviews with administrators from education authority and junior high school headteachers to explore their views about applying a value-added method in Junior high schools' effectiveness evaluations. In addition, I am also seeking your consent for me to collect students' examination results (Junior High School Entrance Examinations, Senior High School Entrance Examinations).

The results of this research will be vital because they will contribute to understanding issues of school effectiveness evaluation and provide research evidence about the application of an alternative method to improve school effectiveness evaluation. The director will receive a report of the research finding once completed and approved by the University. All research participants will remain confidential, and no schools or individuals will be named in any publicly available reports or documents.

Thank you in advance for your support and cooperation.

Yours sincerely,
Ruimin Zhang

Doctoral Student
School of Education
University of Bristol

Request Permission Letter

请求许可信

研究标题：
探讨在中国地方背景下增值评价法作为改善学校效能评估的更为公平的方法的可行性

研究员：
张芮敏， 英国布里斯托大学教育学院博士研究生

此信函是为了请求 W 区教育局批准我进行上述研究。探讨在中国地方背景下增值评价法作为改善学校效能评估的更为公平的方法的可行性。本研究将开展一项问卷调查，涉及 W 区所有公立初中 2021 届九年级学生。该问卷将收集学生及其家庭的基本信息。这些数据将有助于本人进行数据分析，以估算学校对学生中考成绩的影响。该研究还涉及对教育局管理员和初中校长的访谈，以探讨他们对在初中学校效能评估中运用增值评价法的看法。此外，我也请求贵局的许可，能够收集本区初三年级学生的初中入学考试成绩以及中考考试成绩，该数据将有助于本人进行数据分析。

该研究的结果将是至关重要的，因为它们将有助于理解学校效能评估问题，并提供关于运用增值评价法改善学校效能评估的研究证据。研究结果一旦完成并获得大学的批准，将呈报给贵局局长。所有研究参与者将保持机密，任何公开可用的报告或文件中不会透露任何学校或个人的名称。

感谢您的支持和合作。

英国布里斯托大学教育研究生院
博士研究生 张芮敏

Appendix 2 Student Questionnaire and Consent

Dear student

This questionnaire is designed to collect your personal and some of your home background information. We want to know this information to feed into our research that aims to explore the value-added approach's feasibility as an alternative method to improve school effectiveness evaluation. This questionnaire is given to all Grade 9 students in 2021 in public Junior High Schools in the W District. Your participation will be very helpful to the success of this research.

All information you provide voluntarily will be treated as strictly confidential and anonymous-nobody outside the researcher will have access to your detailed responses. Your answers will be combined with others to make totals and averages in which no individual can be identified. Therefore, please feel free to fill out this questionnaire. If you would prefer not to take part, you can contact the researcher and are free to opt-out at any time without giving a reason.

If you need any help, please ask the researcher, Zhang Ruimin (EdD student, ****, ed18369@bristol.ac.uk) or otherwise the supervisor, Sally Thomas (Professor, University of Bristol, S.Thomas@bristol.ac.uk). Please contact * (Education Bureau of W District staff, ***, ***) if you have any complaints about this research.

Thank you for taking the time to fill this questionnaire. I am appreciated with your help!

*** Completing the questionnaire indicates that you consent to participate and have read and agree with the research conditions outlined above.**

1. Current school full name: _____
2. Current school ID: _____
3. Current class teacher's name: _____
4. Student ID: _____
5. Senior High School Entrance Examination Registration Number: _____
Your personal and home background information
6. Gender (please tick " √ " only <u>one</u> box): Male □ Female □
7. Date of birth (please write <u>the number</u> in the blank provided):_____Year_____Month
8. Are you an only child(please tick " √ " only <u>one yes or no</u> box)? No □ Yes □

9. Status of your Hukou (please tick " √ " only one box): Agricultural □ Non-Agricultural □	
10. Are you a transfer student from another district (please tick " √ " only one yes or no box)? No □ Yes □	
11. How many years have you been at this school (please write the number in the blank provided)? _____	
12. Did you take any non-academic curriculum related training **outside of school** in the last three years (e.g., Arts, Programming, football)? (please tick " √ " only one yes or no box) No □ Yes □	
13. Did you take any academic curriculum related training **outside of school** in the last three years (e.g., English, Math)? (please tick " √ " only one yes or no box) No □ Yes □	
14. How many rooms (excluding kitchen and bathroom) are there in your home (please write the number in the blank provided)? _____	
15. Can following items be found in your home? (please tick " √ " only one yes or no box for each item)	
Televisions	No □ Yes □
Computers (desktop, laptop)	No □ Yes □
Cars	No □ Yes □
Smartphones	No □ Yes □
iPad	No □ Yes □
e-book readers	No □ Yes □
Musical instruments	No □ Yes □
16. Do you own yourself the following items? (please tick " √ " only one yes or no box for each item)	
Smartphones	No □ Yes □
iPad	No □ Yes □
Computers (desktop, laptop)	No □ Yes □
E-book readers	No □ Yes □
Room	No □ Yes □
17. How many books are there in your home? (please tick " √ " only one box)	

0 □ 1-10 □ 11-100 □ 101-200 □ 201 or above □
18. What is the highest level of schooling completed by your father? (please tick " √ " only one box) Junior school □ Junior High School□ Senior High School, or Vocational High School □ College □ Bachelor □ Master or above □ Others □
19. What is the highest level of schooling completed by your mother? (please tick " √ " only one box) Junior school □ Junior High School□ Senior High School, or Vocational High School □ College □ Bachelor □ Master or above □ Others □
20. What is your father's main job? (please tick " √ " only one box) Public servants of state-owned enterprises, state organs, and other state units □ Professional and technical personnel (e.g., teacher, doctor, lawyer, accountant) □ Owners or managers of private enterprises, and individual businesses □ Production, Transportation, Equipment operators and related worker or labors □ Employees and service personnel of private enterprises □ Agriculture and Water conservancy labors □ Migrant worker □ Part-time workers □ Other jobs not listed above □ Unemployed □

21. What is your mother's main job? (please tick " √ " only one box)

Public servants of state-owned enterprises, state organs, and other state units □

Professional and technical personnel (e.g., teacher, doctor, lawyer, accountant) □

Owners or managers of private enterprises, and individual businesses □

Production, Transportation, Equipment operators and related worker or labors □

Employees and service personnel of private enterprises □

Agriculture and Water conservancy labors □

Migrant worker □

Part-time workers □

Other jobs not listed above □

Unemployed □

Appendix 3 Interview Participant Consent Form

Research Title: Exploring the feasibility of the value-added approach as a fairer method to measure public junior high school performance in the context of China.

Researcher: Ruimin Zhang, Graduate School of Education, University of Bristol, UK

Supervisor: Professor Sally Thomas, Graduate School of Education, University of Bristol, UK

The aims of the study are to provide new empirical research evidence from the local context of China for exploring the feasibility of the value-added approach as a fair method to measure school effectiveness. Participants in the first phase of research are the Grade 9 students from junior high schools in W District. Participants in the second phase of research are stakeholders of junior high schools. Participants will take part in a one-to-one and face-to-face interview. The interview will last between 30 and 60 minutes and will be audio-recorded.

What You Will Be Asked to Do in the Research: To participate in anonymous interview

Voluntary Participation: Your participation in the study is completely voluntary and you may refuse to answer any question or choose to stop participating at any time. Your decision not to volunteer will not influence your relationship with the researcher or the University of Bristol, either now or in the future.

Withdrawal from the Study: You can stop participating in the study at any time, for any reason, if you decide so. Your decision to stop participating, or to refuse to answer particular questions, will not affect your relationship with the researcher or the University of Bristol. Should you decide to withdraw from the study, all data generated as a consequence of your participation will be destroyed。

Confidentiality: All information you supply during the research will be held confidentially and your name will not appear in any report or publication of the research. Your data will be safely stored in a locked facility and only the researcher will have access to this information.

Questions about the Research: If you have questions about the research in general or about your role in the study, please feel free to contact Ruimin Zhang, EdD Student, at the School of Education, 35 Berkeley Square, Bristol, BS8 1JA, telephone *** or by e-mail (ed18369@bristol.ac.uk). Or otherwise, the supervisor, Sally Thomas (Professor, University of Bristol, S.Thomas@bristol.ac.uk). Please contact * (Education Bureau of W District staff, ***, ***) if you have any complaints about this research.

Consent statement:

I agree to participate in this research, and I am aware that I am free to withdraw at any point without giving a reason by contacting Zhang Ruimin. I understand that my data may not be erased if I do withdraw but will only be used in an anonymized form as part of an aggregated dataset. I understand that the personal data collected from me during the research will be used for the purposes outlined above.

Signature: Date:

Appendix 4 Interview Schedule with Policymakers and Headteachers

Section 1: Overall views on school evaluation and current outcome-based school evaluation practices

1. What is the purpose of school evaluation in W district?
2. What are your views (strengths and weaknesses) about the current practices of public junior high school outcome-based evaluation? Can you provide one or two explicit examples?
3. Do you think school evaluation is helpful to improve school quality? If yes, how can it improve education quality? If not, why this happens?

Section 2: The concept of Value added and Value-added evaluation

1. What is your understanding of the concept of value added?
2. What is your understanding of the value-added evaluation?
3. Do you think who might use value-added measures? For what purpose/s? (for example, school improvement, teacher professional development, parental school choice, accountability, targeting funding)

Section 3: Potentials of implementing value-added evaluation

4. Do you support using value-added approaches to measure school performance? Why/Why not? Can you give examples?
5. Are there any difficulties in collecting student background information for school performance evaluation?
6. Do you have the technicists to conduct a more complicated statistical analysis?
7. Will teachers, parents, and students understand value-added?
8. Are there any other factors may help the usefulness or application of the value-added approach in practice?
9. Are there any other factors may hinder the usefulness or application of the value-added approach in practice?
10. Are there any unintended outcomes of using value-added data? If yes, can you provide one or two explicit examples?

访谈提纲

第一部分：学校评价

1. 就您个人的看法，学校评价的目的是什么？您认为学校评价的目的是基于问责制的需要还是促进学校改进发展，或者两者均兼备？

2. 能否请您告诉我，现阶段您所管理的区/学校内部对学校（初中）教学质量评价或测量有 哪些相关的政策和实践？目前有什么来自于外部（例如上级主管部门、社会）对您所管理的 区/学校的学校（初中）教学质量评价的相关政策和实践？能否请您举例说明？

3. 对于上述的学校（初中）教学质量评价政策和实践，您有什么看法（优点和缺点）？

第二部分：增值评价

1. 在《深化新时代教育评价改革总体方案》中提到的增值评价，您对此是否了解？您是否知道哪些省份或城市已经将增值评价方法应用在学校评价工作中？
2. 您是如何理解“增值”的概念？
3. 您是如何理解“增值评价”方法？
4. 如果将增值评价法应用在学校评价工作中，您认为哪些机构或主体会在增值评价实践中起主导作用？哪些机构或主体会使用增值评价的结果？基于怎样的目的？（例如，学校改进、教师专业发展、学校/教师问责、教育经费分配等）

第三部分：采用增值评价进行学校效能评价的潜在可行性

1. 您是否在本区/学校评价工作中应用过增值评价法?如果有，您对增值评价法的实践有什么看法?如果没有，您是否会考虑在未来的学校评价工作中尝试采用增值评价法？为什么?
2. 在当前的评价实践中，你认为收集学生背景信息的工作有困难吗？如果有，能否举例说明？
3. 当前，对于具有统计分析专业能力的人是否缺乏？
4. 您认为，老师、家长和学生能理解‘增值’的概念吗？
5. 您认为，有什么因素可以促进增值评价方法的实施？
6. 您认为，有什么因素将阻碍增值评价方法的实施？
7. 如果采用增值评价，您认为是否会产生出乎计划的结果？如果有，能否举例说明？

Appendix 5 GSoE Research Ethics Form

Name(s): Zhang Ruimin

Proposed research project: Exploring the feasibility of the value-added approach as an alternative method to measure junior high school performance in the context of China.
Proposed funder(s): N/A
Discussant for the ethics meeting: Jing Zhang
Name of supervisor: Prof. Sally Thomas
Has your supervisor seen this submitted draft of your ethics application? Y

Please include an outline of the project or append a short (1 page) summary:

In 2020, the Chinese government issued an overall plan on educational evaluation reform, which calls for exploring the value-added approach to measure school effectiveness. This quant-qual mixed-method study aims to explore the feasibility of the value-added approach as an alternative method to measure junior high school performance in one district of a city in China.

The purpose of this study's quantitative phase is to find the junior high school performance once the value-added approach is adapted. Ten public junior high schools in this district will be targeted. In terms of Grade 9 students, a total of about 2000 students will be targeted. The primary technique for collecting data will be a questionnaire for Grade 9 students. It is supposed to collect the students' personal and some of their family background information.

In the qualitative phase of this study, the guiding research question is to explore the key factors that seem to help or hinder the development of the value-added approach in school performance evaluation practice. The purposeful sample will be four headteachers from those junior high schools and two school evaluation administrators from the Local Education Authority. They will take part in a 30 to 60 minutes interview by telephone with audio-recording.

Because of the characteristic of multilevel data (students nested within schools), this study will use multilevel models to analyse the quantitative data collected through questionnaires and find the range and extent of school performance after the value-added approach is employed. The qualitative analytic method in this study is thematic analysis. The text obtained through the interviews will be coded and analysed for themes with Nvivo software's help.

Ethical issues discussed and decisions taken as related to:

1. **Researcher access/exit**

Research participants are students (age 15), headteachers, and managers from Junior High schools in one district of China. To recruit and enrol them in the study, the researcher has obtained permission and supports from W District Education Authority. One manager and two staff from the LEA will support the data collection work and help the researcher dispatch invitation letters and consent forms to schools for students' survey. All ten schools have expressed their commitment to take part and

agreed to arrange a student questionnaire. In the end, a report of the research results will be shared with the LEA. There are no financial benefits involved in the research. Invitation letters for students will be dispatched to schools. Students will be given prior information about the research and the opportunity to opt-out by their headteachers.

2. Power and participant relations

The researcher does not work for the education authority or the school, and this is an academic study with no relation to education authority monitor and evaluation. Therefore, all participants will not be at risk of being coerced. A letter of identification from the local education authority will be provided to prove researcher identification. Participants will be given autonomy, and they will have the choice to provide the information. To guarantee this, the researcher will obtain permission from targeted interviewees, and interviewees will read and sign a consent form before the interview.

3. Information given to participants.

- A letter of permission to conduct research from the local education authority
- letter to participants (for students and interviewees)
- Participant Consent Form
- Researcher's and supervisor contact information
- LEA support staff contact information for participants' complaints
- Supervisor's information

4. Participant's right of withdrawal

Students and interviewees have the discretion on whether to participate in the research. It will be made clear in the letter to participants and the consent forms that there will be no consequences if they choose to do so.

5. Informed Consent

According to China's legal requirement, the consent can be acquired from the student with age above 14 (GB/T 35273-2017 Information Technology – Personal Information Security Specification). Apart from this, the staff from the education authority will help provide interviewees' contact information.

Considering that it is a highly centralized education system in China, all targeted junior high schools are directly under the Education Bureau of W District management. Thus, students' consent will be collectively sought at the local education authority and the school's level. They can contact the researcher if they want to opt-out. Apart from this, the student will be informed in the information letter of the questionnaire that completing and returning the questionnaire indicates active opt-in consent. Besides, interviewees should read and sign a consent form before the interview.

6. Complaints procedure

If the complaint concerns suspected research misconduct, the Bristol university procedure should be followed.

Information regarding complaints procedure should be translated in Chinese and given to participants.

7. Safety and well-being of participants/researchers

Student questionnaires will take place in schools in paper form within the regular school day. The time will be arranged by schools. Interviews will take place over the telephone. Survey will be completed confidentially using only numerical school, class, and student IDs.

8. Anonymity/confidentiality

All data collected will be confidential. Any information identifying classrooms, schools, teachers, and pupils will be removed from all paperwork and the unique code will be used instead. Therefore, all research participants will remain confidential, and no schools or individuals will be named in any reports or documents produced based on this study.

9. Data collection

Students' questionnaires in paper form will be given to each school coordinator with the LEA support staff's help, and then the school will arrange students to fill the questionnaires and collect them back to the LEA support staff and researcher.

Interviewees will receive the researcher's interview request and arrange the time to take place the interview through telephone.

10. Data analysis

One step in data analysis is to convert information from the paper questionnaire to digital information. The researcher will conduct this work in the LEA office with the instruction of the LEA. The other work will be the translations for the data from the interview.

11. Data storage

All the paper documents will be stored securely in the LEA office. The digital data will be stored securely on the research data storage facility of the university.

Only the research team members, one manager and two support staff, supervisor, and one researcher's Chinese colleague (peer debriefing for interview data) will access the documents.

12. Data protection (see: http://www.bristol.ac.uk/secretary/data-protection/)

The research adheres to the Data Protection Act 2018, General Data Protection Regulation, GDPR (2018), the Data Protection Guidelines set by the university, and GB/T 35273-2017 Information Technology – Personal Information Security Specification in China.

13. Feedback

In the end, a research report will be shared with the LEA. A meeting with the LEA manager and support staff will be organized after the research to receive the LEA feedback.

14. Responsibilities to colleagues/academic community

15. Reporting of research

The results may be used for a local education authority funded project, and the consent from the local education authority and participant will be obtained.

If you feel you need to discuss any issue further, or to highlight difficulties, please contact the GSoE's ethics co-ordinators who will suggest possible ways forward.

Signed: Ruimin Zhang (Researcher) Signed: Jing Zhang (Discussant)

Date: 08/02/21

Appendix 6 Model Equations

SHSEE Total: Raw model

$ZSHSEEtotal_{ijk} \sim N(XB, \Omega)$

$ZSHSEEtotal_{ijk} = \beta_{0ijk}cons$

$\beta_{0ijk} = -0.144(0.151) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.146(0.103)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.257(0.064)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.547(0.020)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3708.235(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Total: VA model

$ZSHSEEtotal_{ijk} \sim N(XB, \Omega)$

$ZSHSEEtotal_{ijk} = \beta_{0ijk}cons + 0.246(0.021)ZJHSEEchi_{ijk} + 0.311(0.021)ZJHSEEmath_{ijk} + 0.099(0.019)ZJHSEEeng_{ijk}$

$\beta_{0ijk} = -0.117(0.099) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.072(0.045)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.078(0.021)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.320(0.011)\right]$

*-2*loglikelihood(IGLS Deviance)* = 2824.737(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Total: CVA-1 model

$ZSHSEEtotal_{ijk} \sim N(XB, \Omega)$

$ZSHSEEtotal_{ijk} = \beta_{0ijk}cons + 0.222(0.021)ZJHSEEchi_{ijk} + 0.307(0.021)ZJHSEEmath_{ijk} + 0.079(0.018)ZJHSEEeng_{ijk} + -0.149(0.029)boy_{ijk} + 0.170(0.030)yes_{ijk} + 0.174(0.070)11\text{-}100\ books_{ijk} + 0.286(0.075)101\text{-}200\ books_{ijk} + 0.317(0.082)201\ or\ above\ books_{ijk}$

$\beta_{0ijk} = -0.328(0.115) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.059(0.038)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.070(0.019)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.302(0.011)\right]$

*-2*loglikelihood(IGLS Deviance)* = 2733.726(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Total: CVA-2 model

$ZSHSEEtotal_{ijk} \sim N(XB, \Omega)$

$ZSHSEEtotal_{ijk} = \beta_{0ijk}cons + 0.218(0.021)ZJHSEEchi_{ijk} + 0.301(0.021)ZJHSEEmath_{ijk} + 0.073(0.018)ZJHSEEeng_{ijk} + -0.149(0.029)boy_{ijk} + 0.166(0.030)yes_{ijk} + 0.167(0.069)11\text{-}100\ books_{ijk} + 0.275(0.075)101\text{-}200\ books_{ijk} + 0.307(0.082)201\ or\ above\ books_{ijk} + 0.254(0.036)ZCpriormeantotal_{jk}$

$\beta_{0ijk} = -0.296(0.107) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.055(0.030)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.026(0.008)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.302(0.011)\right]$

*-2*loglikelihood(IGLS Deviance)* = 2701.993(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Chinese: Raw model

$ZSHSEEchi_{ijk} \sim N(XB, \Omega)$

$ZSHSEEchi_{ijk} = \beta_{0ijk}cons$

$\beta_{0ijk} = -0.142(0.142) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.137(0.092)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.196(0.051)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.647(0.023)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3958.427(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Chinese: VA model

$ZSHSEEchi_{ijk} \sim N(XB, \Omega)$

$ZSHSEEchi_{ijk} = \beta_{0ijk}cons + 0.409(0.024)ZJHSEEchi_{ijk} + 0.221(0.023)ZJHSEEmath_{ijk} + 0.034(0.021)ZJHSEEeng_{ijk}$

$\beta_{0ijk} = -0.107(0.089) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.060(0.036)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.051(0.015)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.394(0.014)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3135.517(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Chinese：CVA-1 model

$ZSHSEEchi_{ijk} \sim N(XB, \Omega)$

$ZSHSEEchi_{ijk} = \beta_{0ijk}cons + 0.371(0.023)ZJHSEEchi_{ijk} + 0.228(0.023)ZJHSEEmath_{ijk} + 0.012(0.020)ZJHSEEeng_{ijk} + -0.245(0.032)boy_{ijk} + 0.090(0.033)yes_{ijk} + 0.233(0.077)11\text{-}100\ books_{ijk} + 0.380(0.083)101\text{-}200\ books_{ijk} + 0.383(0.090)201\ or\ above\ books_{ijk}$

$\beta_{0ijk} = -0.302(0.113) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.050(0.031)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.046(0.013)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.369(0.013)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3030.302(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Chinese: CVA-2 model

$ZSHSEEchi_{ijk} \sim N(XB, \Omega)$

$ZSHSEEchi_{ijk} = \beta_{0ijk}cons + 0.366(0.023)ZJHSEEchi_{ijk} + 0.223(0.023)ZJHSEEmath_{ijk} + 0.006(0.020)ZJHSEEeng_{ijk} + -0.244(0.032)boy_{ijk} + 0.087(0.033)yes_{ijk} + 0.228(0.077)11\text{-}100\ books_{ijk} + 0.372(0.083)101\text{-}200\ books_{ijk} + 0.376(0.090)201\ or\ above\ books_{ijk} + 0.151(0.039)ZCpriormeantotal_{jk}$

$\beta_{0ijk} = -0.277(0.105) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.036(0.023)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.032(0.010)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.370(0.013)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3017.562(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Mathematics: Raw model

$ZSHSEEmath_{ijk} \sim N(XB, \Omega)$

$ZSHSEEmath_{ijk} = \beta_{0ijk}cons$

$\beta_{0ijk} = -0.118(0.143) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.127(0.091)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.235(0.059)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.590(0.021)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3821.235(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Mathematics: VA model

$ZSHSEEmath_{ijk} \sim N(XB, \Omega)$

$ZSHSEEmath_{ijk} = \beta_{0ijk}cons + 0.108(0.024)ZJHSEEchi_{ijk} + 0.427(0.023)ZJHSEEmath_{ijk} + 0.050(0.021)ZJHSEEeng_{ijk}$

$\beta_{0ijk} = -0.088(0.098) + v_{0k} + u_{0jk} + e_{0ijk}$

$\begin{bmatrix} v_{0k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.067(0.044) \end{bmatrix}$

$\begin{bmatrix} u_{0jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.085(0.023) \end{bmatrix}$

$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.388(0.014) \end{bmatrix}$

*-2*loglikelihood(IGLS Deviance)* = 3128.581(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Mathematics: CVA-1 model

$ZSHSEEmath_{ijk} \sim N(XB, \Omega)$

$ZSHSEEmath_{ijk} = \beta_{0ijk}cons + 0.104(0.024)ZJHSEEchi_{ijk} + 0.412(0.023)ZJHSEEmath_{ijk} + 0.039(0.021)ZJHSEEeng_{ijk} + -0.001(0.032)boy_{ijk} + 0.153(0.034)yes_{ijk} + 0.135(0.078)11\text{-}100\ books_{ijk}$
$0.228(0.084)101\text{-}200\ books_{ijk} + 0.264(0.092)201\ or\ above\ books_{ijk}$

$\beta_{0ijk} = -0.321(0.120) + v_{0k} + u_{0jk} + e_{0ijk}$

$\begin{bmatrix} v_{0k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.054(0.037) \end{bmatrix}$

$\begin{bmatrix} u_{0jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.080(0.021) \end{bmatrix}$

$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.380(0.014) \end{bmatrix}$

*-2*loglikelihood(IGLS Deviance)* = 3092.950(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE Mathematics: CVA-2 model

$ZSHSEEmath_{ijk} \sim N(XB, \Omega)$

$ZSHSEEmath_{ijk} = \beta_{0ijk}cons + 0.099(0.024)ZJHSEEchi_{ijk} + 0.407(0.023)ZJHSEEmath_{ijk} + 0.034(0.021)ZJHSEEeng_{ijk} + -0.000(0.032)boy_{ijk} + 0.147(0.034)yes_{ijk} + 0.128(0.078)11\text{-}100\ books_{ijk}$
$0.217(0.084)101\text{-}200\ books_{ijk} + 0.254(0.092)201\ or\ above\ books_{ijk} + 0.253(0.041)ZCpriormeantotal_{jk}$

$\beta_{0ijk} = -0.286(0.116) + v_{0k} + u_{0jk} + e_{0ijk}$

$\begin{bmatrix} v_{0k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.058(0.033) \end{bmatrix}$

$\begin{bmatrix} u_{0jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.034(0.011) \end{bmatrix}$

$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.380(0.014) \end{bmatrix}$

*-2*loglikelihood(IGLS Deviance)* = 3066.889(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE English: Raw model

$ZSHSEEeng_{ijk} \sim N(XB, \Omega)$

$ZSHSEEeng_{ijk} = \beta_{0ijk}cons$

$\beta_{0ijk} = -0.137(0.143) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.122(0.091)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.258(0.065)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.578(0.021)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3792.299(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE English: VA model

$ZSHSEEeng_{ijk} \sim N(XB, \Omega)$

$ZSHSEEeng_{ijk} = \beta_{0ijk}cons + 0.213(0.024)ZJHSEEchi_{ijk} + 0.171(0.024)ZJHSEEmath_{ijk} + 0.192(0.021)ZJHSEEeng_{ijk}$

$\beta_{0ijk} = -0.117(0.099) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.064(0.044)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.101(0.027)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.415(0.015)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3240.165(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE English: CVA-1 model

$ZSHSEEeng_{ijk} \sim N(XB, \Omega)$

$ZSHSEEeng_{ijk} = \beta_{0ijk}cons + 0.184(0.024)ZJHSEEchi_{ijk} + 0.171(0.023)ZJHSEEmath_{ijk} + 0.166(0.021)ZJHSEEeng_{ijk} + -0.208(0.033)boy_{ijk} + 0.220(0.034)yes_{ijk} + 0.128(0.079)11\text{-}100\ books_{ijk} - 0.210(0.085)101\text{-}200\ books_{ijk} + 0.252(0.093)201\ or\ above\ books_{ijk}$

$\beta_{0ijk} = -0.269(0.122) + v_{0k} + u_{0jk} + e_{0ijk}$

$\left[v_{0k}\right] \sim N(0, \Omega_v) : \Omega_v = \left[0.055(0.039)\right]$

$\left[u_{0jk}\right] \sim N(0, \Omega_u) : \Omega_u = \left[0.090(0.024)\right]$

$\left[e_{0ijk}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.390(0.014)\right]$

*-2*loglikelihood(IGLS Deviance)* = 3139.486(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

SHSEE English: CVA-2 model

$ZSHSEEeng_{ijk} \sim N(XB, \Omega)$

$ZSHSEEeng_{ijk} = \beta_{0ijk}cons + 0.179(0.024)ZJHSEEchi_{ijk} + 0.165(0.023)ZJHSEEmath_{ijk} + 0.160(0.021)ZJHSEEeng_{ijk} + -0.208(0.033)boy_{ijk} + 0.216(0.034)yes_{ijk} + 0.119(0.079)11\text{-}100\ books_{ijk} + 0.196(0.085)101\text{-}200\ books_{ijk} + 0.240(0.093)201\ or\ above\ books_{ijk} + 0.273(0.043)ZCpriormeantotal_{jk}$

$\beta_{0ijk} = -0.239(0.113) + v_{0k} + u_{0jk} + e_{0ijk}$

$\begin{bmatrix} v_{0k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.050(0.030) \end{bmatrix}$

$\begin{bmatrix} u_{0jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.039(0.012) \end{bmatrix}$

$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.390(0.014) \end{bmatrix}$

*-2*loglikelihood(IGLS Deviance)* = 3110.915(1596 of 1596 cases in use)

UNITS:

SCHOOL: 11 (of 11) in use

CLASS: 46 (of 46) in use

Appendix 7 Model Comparisons

SHSEE Total score

	RM SHSEE TOTAL	S.E.	z-ratio	p-value	VA SHSEE TOTAL	S.E.	z-ratio	p-value	CVA-1 SHSEE TOTAL	S.E.	z-ratio	p-value	CVA-2 SHSEE TOTAL	S.E.	z-ratio	p-value
Response	ZSHSEEtotal				ZSHSEEtotal				ZSHSEEtotal				ZSHSEEtotal			
Fixed Part																
cons	-0.145	0.151	-0.956	0.339	-0.390	0.112	-3.484	0.000	-0.546	0.125	-4.359	0.000	-0.498	0.118	-4.221	0.000
ZJHSEEchi					0.246	0.021	11.461	0.000	0.222	0.021	10.484	0.000	0.218	0.021	10.320	0.000
ZJHSEEmath					0.311	0.021	14.907	0.000	0.307	0.021	14.917	0.000	0.301	0.021	14.662	0.000
JHSEEeng					0.004	0.001	5.281	0.000	0.004	0.001	4.293	0.000	0.003	0.001	3.982	0.000
boy									-0.149	0.029	-5.189	0.000	-0.149	0.029	-5.175	0.000
yes									0.170	0.030	5.601	0.000	0.166	0.030	5.499	0.000
11-100 books									0.174	0.070	2.503	0.012	0.167	0.069	2.402	0.016
101-200 books									0.286	0.075	3.820	0.000	0.275	0.075	3.686	0.000
201 or above books									0.317	0.082	3.869	0.000	0.307	0.082	3.762	0.000
ZCpriormeantotal													0.254	0.036	6.951	0.000
Random Part																
Level: SCHOOL																
Var(cons)	0.146	0.102			0.072	0.045			0.059	0.038			0.055	0.030		
Level: CLASS																
Var(cons)	0.256	0.064			0.078	0.021			0.070	0.019			0.026	0.008		
Level: STUDENT																
Var(cons)	0.547	0.020			0.320	0.011			0.302	0.011			0.302	0.011		
Units: SCHOOL	11				11				11				11			
Units: CLASS	46				46				46				46			
Units: STUDENT	1596				1596				1596				1596			
Estimation:	IGLS				IGLS				IGLS				IGLS			
-2*loglikelihood:	3708.235				2824.737				2733.726				2701.993			

SHSEE Chinese

	RM SHSEE CHINESE	S.E.	z-ratio	p-value	VA SHSEE CHINESE	S.E.	z-ratio	p-value	CVA-1 SHSEE CHINESE	S.E.	z-ratio	p-value	CVA-2 SHSEE CHINESE	S.E.	z-ratio	p-value
Response	ZSHSEEchi				ZSHSEEchi				ZSHSEEchi				ZSHSEEchi			
Fixed Part																
cons	-0.142	0.142	-0.998	0.318	-0.107	0.089	-1.205	0.228	-0.302	0.113	-2.684	0.007	-0.277	0.105	-2.653	0.008
ZJHSEEchi					0.409	0.024	17.246	0.000	0.371	0.023	15.871	0.000	0.366	0.023	15.674	0.000
ZJHSEEmath					0.221	0.023	9.571	0.000	0.228	0.023	10.052	0.000	0.223	0.023	9.820	0.000
ZJHSEEeng					0.034	0.021	1.630	0.103	0.012	0.020	0.584	0.559	0.006	0.020	0.292	0.770
boy									-0.245	0.032	-7.685	0.000	-0.244	0.032	-7.667	0.000
yes									0.090	0.033	2.687	0.007	0.087	0.033	2.600	0.009
11-100 books									0.233	0.077	3.034	0.002	0.228	0.077	2.971	0.003
101-200 books									0.380	0.083	4.602	0.000	0.372	0.083	4.508	0.000
201 or above books									0.383	0.090	4.234	0.000	0.376	0.090	4.160	0.000
ZCpriormeantotal													0.151	0.039	3.832	0.000
Random Part																
Level: SCHOOL																
Var(cons)	0.137	0.092			0.060	0.036			0.050	0.031			0.036	0.023		
Level: CLASS																
Var(cons)	0.196	0.051			0.051	0.015			0.046	0.013			0.032	0.010		
Level: STUDENT																
Var(cons)	0.647	0.023			0.394	0.014			0.369	0.013			0.370	0.013		
Units: SCHOOL	11				11				11				11			
Units: CLASS	46				46				46				46			
Units: STUDENT	1596				1596				1596				1596			
Estimation:	IGLS				IGLS				IGLS				IGLS			
-2*loglikelihood:	3958.427				3135.517				3030.302				3017.562			

SHSEE MATHEMATICS

	RM SHSEE MATH	S.E.	z-ratio	p-value	VA SHSEE MATH	S.E.	z-ratio	p-value	CVA-1 SHSEE MATH	S.E.	z-ratio	p-value	CVA-2 SHSEE MATH	S.E.	z-ratio	p-value
Response	ZSHSEEmath				ZSHSEEmath				ZSHSEEmath				ZSHSEEmath			
Fixed Part																
cons	-0.118	0.143	-0.826	0.409	-0.088	0.098	-0.896	0.370	-0.321	0.120	-2.673	0.008	-0.286	0.116	-2.477	0.013
ZJHSEEchi					0.108	0.024	4.561	0.000	0.104	0.024	4.374	0.000	0.099	0.024	4.172	0.000
ZJHSEEmath					0.427	0.023	18.597	0.000	0.412	0.023	17.889	0.000	0.407	0.023	17.648	0.000
ZJHSEEeng					0.050	0.021	2.406	0.016	0.039	0.021	1.910	0.056	0.034	0.021	1.631	0.103
boy									-0.001	0.032	-0.036	0.971	-0.000	0.032	-0.008	0.993
yes									0.153	0.034	4.510	0.000	0.147	0.034	4.345	0.000
11-100 books									0.135	0.078	1.728	0.084	0.128	0.078	1.648	0.099
101-200 books									0.228	0.084	2.712	0.007	0.217	0.084	2.587	0.010
201 or above books									0.264	0.092	2.872	0.004	0.254	0.092	2.768	0.006
ZCpriormeantotal													0.253	0.041	6.116	0.000
Random Part																
Level: SCHOOL																
Var(cons)	0.127	0.092			0.067	0.044			0.054	0.037			0.058	0.033		
Level: CLASS																
Var(cons)	0.235	0.059			0.085	0.023			0.080	0.021			0.034	0.011		
Level: STUDENT																
Var(cons)	0.590	0.021			0.388	0.014			0.380	0.014			0.380	0.014		
Units: SCHOOL	11				11				11				11			
Units: CLASS	46				46				46				46			
Units: STUDENT	1596				1596				1596				1596			
Estimation:	IGLS				IGLS				IGLS				IGLS			
-2*loglikelihood:	3821.235				3128.581				3092.950				3066.889			

SHSEE ENGLISH

	RM SHSEE ENG	S.E.	z-ratio	p-value	VA SHSEE ENG	S.E.	z-ratio	p-value	CVA-1 SHSEE ENG	S.E.	z-ratio	p-value	CVA-2 SHSEE ENG	S.E.	z-ratio	p-value
Response	ZSHSEEeng				ZSHSEEeng				ZSHSEEeng				ZSHSEEeng			
Fixed Part																
cons	-0.137	0.143	-0.955	0.339	-0.117	0.099	-1.174	0.240	-0.269	0.122	-2.198	0.028	-0.239	0.113	-2.109	0.035
ZJHSEEchi					0.213	0.024	8.731	0.000	0.184	0.024	7.657	0.000	0.179	0.024	7.452	0.000
ZJHSEEmath					0.171	0.024	7.218	0.000	0.171	0.023	7.327	0.000	0.165	0.023	7.078	0.000
ZJHSEEeng					0.192	0.021	8.945	0.000	0.166	0.021	7.932	0.000	0.160	0.021	7.657	0.000
boy									-0.208	0.033	-6.359	0.000	-0.208	0.033	-6.364	0.000
yes									0.220	0.034	6.377	0.000	0.216	0.034	6.283	0.000
11-100 books									0.128	0.079	1.621	0.105	0.119	0.079	1.506	0.132
101-200 books									0.210	0.085	2.467	0.014	0.196	0.085	2.307	0.021
201 or above books									0.252	0.093	2.707	0.007	0.240	0.093	2.584	0.010
ZCpriormeantotal													0.273	0.043	6.361	0.000
Random Part																
Level: SCHOOL																
Var(cons)	0.122	0.091			0.064	0.045			0.055	0.039			0.049	0.030		
Level: CLASS																
Var(cons)	0.258	0.065			0.101	0.027			0.090	0.024			0.039	0.012		
Level: STUDENT																
Var(cons)	0.578	0.021			0.415	0.015			0.390	0.014			0.390	0.014		
Units: SCHOOL	11				11				11				11			
Units: CLASS	46				46				46				46			
Units: STUDENT	1596				1596				1596				1596			
Estimation:	IGLS				IGLS				IGLS				IGLS			
-2*loglikelihood:	3792.299				3240.165				3139.486				3110.915			

Appendix 8 Univariable Models for Selected Explanatory Variables: Sample Results

SHSEE Total score	
Hukou (place of household registration)	$ZSHSEEtotal_i \sim N(XB, \Omega)$ $ZSHSEEtotal_i = \beta_{0i}cons + -0.698(0.050)agriculture_i$ $\beta_{0i} = 0.470(0.041) + e_{0i}$ $[e_{0i}] \sim N(0, \Omega_e) : \Omega_e = [0.892(0.032)]$ *-2*loglikelihood(IGLS Deviance)* = 4347.061(1596 of 1596 cases in use)
Onlychild (whether the only child of the family)	$ZSHSEEtotal_i \sim N(XB, \Omega)$ $ZSHSEEtotal_i = \beta_{0i}cons + 0.061(0.060)yes_i$ $\beta_{0i} = -0.014(0.028) + e_{0i}$ $[e_{0i}] \sim N(0, \Omega_e) : \Omega_e = [0.999(0.035)]$ *-2*loglikelihood(IGLS Deviance)* = 4527.192(1596 of 1596 cases in use)
Nonactutoring (whether receive non-academic tutoring outside of school)	$ZSHSEEtotal_i \sim N(XB, \Omega)$ $ZSHSEEtotal_i = \beta_{0i}cons + 0.198(0.060)yes_i$ $\beta_{0i} = -0.045(0.028) + e_{0i}$ $[e_{0i}] \sim N(0, \Omega_e) : \Omega_e = [0.992(0.035)]$ *-2*loglikelihood(IGLS Deviance)* = 4517.209(1596 of 1596 cases in use)
SHSEE Chinese	
Hukou (place of household registration)	$ZSHSEEchi_i \sim N(XB, \Omega)$ $ZSHSEEchi_i = \beta_{0i}cons + -0.609(0.051)agriculture_i$ $\beta_{0i} = 0.410(0.042) + e_{0i}$ $[e_{0i}] \sim N(0, \Omega_e) : \Omega_e = [0.918(0.033)]$ *-2*loglikelihood(IGLS Deviance)* = 4392.286(1596 of 1596 cases in use)
Onlychild (whether the only child of the family)	$ZSHSEEchi_i \sim N(XB, \Omega)$ $ZSHSEEchi_i = \beta_{0i}cons + 0.091(0.060)yes_i$ $\beta_{0i} = -0.021(0.028) + e_{0i}$ $[e_{0i}] \sim N(0, \Omega_e) : \Omega_e = [0.998(0.035)]$ *-2*loglikelihood(IGLS Deviance)* = 4525.901(1596 of 1596 cases in use)

Nonactutoring (whether receive non-academic tutoring outside of school)	$ZSHSEEchi_i \sim N(XB, \Omega)$ $ZSHSEEchi_i = \beta_{0i}cons + 0.120(0.060)yes_i$ $\beta_{0i} = -0.027(0.028) + e_{0i}$ $\left[e_{0i}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.997(0.035)\right]$ *-2*loglikelihood(IGLS Deviance)* = 4524.219(1596 of 1596 cases in use)
SHSEE Mathematics	
Hukou (place of household registration)	$ZSHSEEmath_i \sim N(XB, \Omega)$ $ZSHSEEmath_i = \beta_{0i}cons + -0.627(0.051)agriculture_i$ $\beta_{0i} = 0.422(0.042) + e_{0i}$ $\left[e_{0i}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.913(0.032)\right]$ *-2*loglikelihood(IGLS Deviance)* = 4383.907(1596 of 1596 cases in use)
Onlychild (whether the only child of the family)	$ZSHSEEmath_i \sim N(XB, \Omega)$ $ZSHSEEmath_i = \beta_{0i}cons + 0.018(0.060)yes_i$ $\beta_{0i} = -0.004(0.028) + e_{0i}$ $\left[e_{0i}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.999(0.035)\right]$ *-2*loglikelihood(IGLS Deviance)* = 4528.156(1596 of 1596 cases in use)
Nonactutoring (whether receive non-academic tutoring outside of school)	$ZSHSEEmath_i \sim N(XB, \Omega)$ $ZSHSEEmath_i = \beta_{0i}cons + 0.212(0.060)yes_i$ $\beta_{0i} = -0.048(0.028) + e_{0i}$ $\left[e_{0i}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.992(0.035)\right]$ *-2*loglikelihood(IGLS Deviance)* = 4515.669(1596 of 1596 cases in use)
SHSEE English	
Hukou (place of household registration)	$ZSHSEEeng_i \sim N(XB, \Omega)$ $ZSHSEEeng_i = \beta_{0i}cons + -0.683(0.051)agriculture_i$ $\beta_{0i} = 0.460(0.041) + e_{0i}$ $\left[e_{0i}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.897(0.032)\right]$ *-2*loglikelihood(IGLS Deviance)* = 4355.117(1596 of 1596 cases in use)
Onlychild (whether the only child of the family)	$ZSHSEEeng_i \sim N(XB, \Omega)$ $ZSHSEEeng_i = \beta_{0i}cons + 0.072(0.060)yes_i$ $\beta_{0i} = -0.016(0.028) + e_{0i}$ $\left[e_{0i}\right] \sim N(0, \Omega_e) : \Omega_e = \left[0.998(0.035)\right]$ *-2*loglikelihood(IGLS Deviance)* = 4526.803(1596 of 1596 cases in use)

Nonactutoring (whether receive non-academic tutoring outside of school)	$ZSHSEEeng_i \sim N(XB, \Omega)$ $ZSHSEEeng_i = \beta_{0i}cons + 0.199(0.060)yes_i$ $\beta_{0i} = -0.045(0.028) + e_{0i}$ $\begin{bmatrix} e_{0i} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.992(0.035) \end{bmatrix}$ *-2*loglikelihood(IGLS Deviance)* = 4517.130(1596 of 1596 cases in use)

Explanatory variables tested individually and in combination (significant at P=0.05 level)

SHSEE Total score	SHSEE Chinese	SHSEE Mathematics	SHSEE English
Gender *	Gender *	Gender *	Gender *
Onlychild	Onlychild	Onlychild	Onlychild
Hukou *	Hukou	Hukou	Hukou
TransferD (whether a transfer student from another district)	TransferD	TransferD	TransferD
TransferS (whether a transfer student from another school)	TransferS	TransferS	TransferS
Nonactutoring	Nonactutoring	Nonactutoring	Nonactutoring
Actutoring *	Actutoring *	Actutoring *	Actutoring *
Nroom (number of rooms in the home)	Nroom	Nroom	Nroom
Items in the home			
Televisions	Televisions	Televisions	Televisions
Computers *	Computers *	Computers	Computers
Cars *	Cars	Cars	Cars
Smartphones *	Smartphones	Smartphones	Smartphones
iPad	iPad	iPad *	iPad
e-book readers	e-book readers	e-book readers	e-book readers
Musical instruments	Musical instruments	Musical instruments	Musical instruments
Items owned by student			
Ssmartphones *	Ssmartphones	Ssmartphones *	Ssmartphones
SiPad	SiPad	SiPad	SiPad
Scomputers	Scomputers	Scomputers	Scomputers
Se-book readers	Se-book readers	Se-book readers	Se-book readers
Sroom	Sroom	Sroom	Sroom
Books * (number of books in the home)	Books *	Books *	Books *
Father's education	Father's education	Father's education	Father's education
Mother's education	Mother's education	Mother's education	Mother's education
Father's occupation *	Father's occupation	Father's occupation *	Father's occupation
Mother's occupation	Mother's occupation	Mother's occupation	Mother's occupation

*Note: Variables marked with * remained significant at the 0.05 level when tested jointly with other variables*

Appendix 9 Themes and Codes

<table>
<tr><td colspan="4">Sub RQ2-1: What are interviewee perceptions on the purpose of current junior high school academic performance evaluation?</td></tr>
<tr><td>Theme</td><td>Codes</td><td>Sub-codes</td><td>Example Quotes</td></tr>
<tr><td rowspan="2">Purpose of School Evaluation</td><td>Meeting School Inspection Requirements</td><td>1.Ensuring Standards and Expectations
2.Monitoring High-Quality School Education</td><td>"I think the primary purpose of school evaluation is to ensure that schools are meeting certain standards and expectations…… In the context of China, one aspect of school inspection involves using student academic outcome data to assess whether schools have achieved pre-established goals or standards. ” (PM2)
“I think the purpose of school evaluation is to meet one of the school inspection requirements. It serves as a critical component in monitoring high-quality of school education. ” (HT2)</td></tr>
<tr><td>Supporting School improvement</td><td>1. Providing Feedback and Identifying Strengths and Weaknesses
2. Informing Policy Monitoring and Goal Establishment
3.Sharing Successful Experiences and Providing Support</td><td>“I think the purpose of school evaluation is to provide helpful feedback information to schools with identification of their strengths and weaknesses and provide direction for school improvement. For example, if evaluation results show that students' academic performance is poor, the school can focus on curriculum design or teacher professional development to address the issue. ”(HT1)
“evaluation results can help policymakers to monitor policy implementation and establish meaningful goals for schools” (PM2)
“high-performing schools could share successful experiences with low-performing schools and provide support for their improvement”. (PM1)</td></tr>
<tr><td colspan="4">Sub RQ2-2: What are interviewee perceptions of the benefits and disadvantages of the raw attainment school performance measures in the local context of this study?</td></tr>
</table>

Benefits and Disadvantages of the Unadjusted Raw attainment-based School Performance Measures	Advantages of Raw Attainment Measures	1.Simple and Transparency 2.Enabling the simplicity of comparison 3.Guiding Policymaking and Goal setting 4.Stimulating Improvement	*"There have been no changes in school academic performance evaluation practices over the past few years. Student results in SHSEE are a straightforward indicator of school academic performance. Education authorities, schools, and teachers can easily compare and understand this indicator for showing a school's success."* (PM2) *"Receiving the feedback of our students' performance in SHSEE, we are able to identify areas that doing well and that need improvement. This information helps us make informed decisions, develop targeted interventions, and establish meaningful goals to drive educational improvement.. "* (HT1) *"Recognitions received from the LEA motivate us to continue academic success. Put another way, high-performing results provide information that we are teaching appropriately to enable students to achieve better attainment."* (HT3)
	Disadvantages of raw attainment measures	1.Influence of Student and School Context 2.Limitation in Capturing External Factors, Effort of Teachers and Students 3.Focus on Aggregate Data 4.Bias and Narrow Focus	*"This school is in a rural area, and the overall student entrance level is lower than average. This year, the measurement of results for our school were at the bottom. However, I know our teachers and students have worked hard. To some extent, this frustrated our teachers. " (HT1)* *"It seems that raw measurements may not always fully capture the impact of these factors and may not always accurately reflect the hard work that teachers and students put into their work." (HT2)* *"I think it is problematic to talk about school outcomes without caring about individual students. For example, did girls do equally well as boys? Did students from rural areas perform with higher or lower mark than those from urban areas?" (PM3)* *"Teachers tend to focus on two groups of students to ensure a good classification performance. One group is students near the pass standard, and the other group is students with a good chance of achieving high exam marks. "* (HT1)

Sub RQ2-3: What are interviewee perceptions on the concept of VAMs?			
The Concept of Value-Added	1. Different Perspectives on the Concept of VAMs 2.The Importance of Considering Baseline Assessment 3.The Need to Consider Contextual Factors		*"calculating the difference between a student's raw attainment in the entrance examinations and the school-leaving examinations would indicate a school's added value."* (HT1) *"Measuring a school's added value could be considered as comparing observed student marks to their expected marks in leaving examinations".* (HT2) *"The school's contribution to student progress in academic attainment".* (PM2) "*In order to accurately assess a school's performance, I think it's essential to take into account students' starting points and the progress they make over time*." (PM2) *"I think it's important to consider contextual factors such as school size, status, type, and location. For example, a small rural school may face different challenges than a large urban school with a diverse student population. Different school type may have various educational resources. Therefore, I think we need to take into account the context in which the school operates. This may help us to make a fair evaluation of school performance."* (PM3)
Sub RQ2-4: To what extent (strong or weak) do interviewees have a motivation to implement VAMs in practice?			
Motivations to Implement VAMs in Practice	Strong Motivation to Employ VAMs	1. Desire for Fairer School Comparison 2. Recognition of Limitations in Raw Attainment Measures 3. The Need for Comprehensive Assessment of School Impact 4. The Drive to Identify Areas for Improvement	*"We usually employ student examination results as one of the means to monitor school performance. Since value-added approaches can enable a fairer school comparison than those based on students' raw examination marks at a single point, we are interested in applying it."* (PM1) *"We recognize the limitations of using raw examination results to measure school performance. As I mentioned before, unadjusted raw attainment measures fail to consider the influence of school level contextual factors. I am open to the idea of VAMs and would like to learn more about how it could be implemented in our context. "* (PM3) *"Although raw examination results were used by*

			the school inspection as one of the indicators of school academic performance, I am still thinking about the approach that can help us better understand our school's impact on student learning and identify areas where we can make improvements. Thus, I am interested to look at how VAMs can provide a comprehensive picture of our school's performance." (HT2)
	Weak Motivation to Employ VAMs	1. Focus on High-Stakes Examinations 2. Lack of Decision-Making Power 3. Perceived Limited Relevance or Utility	*"I agree with that there are the potential benefits of VAMs, but the reality is that schools are under a lot of pressure to ensure students achieve target marks in high-stakes exams. Even if we use VAMs, the focus is still on examination outcomes, and our pressure from achieving high marks in high-stakes examinations is not relieved."* (HT3) *"I understand the potential benefits of using VAMs to evaluate school performance, but the decision-making power for implementing these approaches lies with the local education authority. As a headteacher, I may not have the authority to implement such methods, which can limit my motivation to pursue it further."* (HT1) *"I don't think VAMs is something that we would use extensively at our school. I see it more as a tool that external inspectors might use to evaluate our school. Therefore, we pay less attention or interests in investigating this approach."* (HT3)
Sub RQ2-5: What are interviewee perceptions of the factors that may enhance or hinder the implementation of the value-added approach in junior high school effectiveness evaluation in the context of this study?			
Factors Influence the Implementation of VAMs	Factors Supporting the Implementation of VAMs	1. Policy Support and Attention 2. Autonomy and Interest of Higher Level of Education Authorities 3. Learning from Successful Implementations 4. Improved Data Systems	*"The potential motivation of exploring the implementation of VAMs at the local level was primarily associated with the new policy document on education evaluation reform. It raised the attention of provincial education authorities to investigate the implementation plans for VAMs in the local context." (PM3)* *"We are the grassroots educational management organizations. Compared to provincial-level education departments, our autonomy in conducting education reform is limited. Thus, it would be better that the higher level of education authorities has the interest in VAMs. " (PM3)* *"We are planning to visit a city in northeast*

			China to learn about the implementation of the value-added evaluation approach set up in the city to promote internal school accountability. I think it is good that the value-added evaluation approach has been used to measure school performance in some regions. It provides a valuable reference for our practice in the future." (PM1) *"Improvement of the data system enable our students' information to be managed in a more effective way. For example, students' enrolment and examination registration are done online, which facilitates the information query. As I know, the availability of the student unique ID enables you to collect and matching data requirement for your study. "(PM2)*
	Factors Hindering the Implementation of VAMs	1. Potential Challenges of Other Education Policies and Priorities Requirement 2. Data Availability and Quality 3. Operational Difficulties 4. Challenges in Interpretation and Understanding of Results 5. Potential Conflict Regarding the Usage of Unadjusted Raw Attainment Measures Versus VAMs Results	*"The "double reduction education policy" may potentially affect the implementation of VAMs. This policy aims to reduce the burden of excessive homework and private tutoring for students, thereby bringing the focus on improving classroom teaching and learning. Consequently, formative assessment at the class level has received greater emphasis, and enhancing formative assessment at the classroom level may impact resource allocation for supporting VAMs."* (PM1) *"To ease students' burden of excessive homework and off-campus tutoring, the government call for reducing the number and frequency of large-scale and standards-based examinations during the phase of primary and junior high schooling. Therefore, the student prior attainment may be unavailable in the future."* (PM2) *".... Instead, external bodies are responsible for collecting this information for various purposes, including education quality monitoring and research projects. They may face the challenges in accessing the data. "* (HT2) *"even though the information collected is anonymous, parents may still have concerns or complaints about these activities."* (HT3) *"I think one of the main challenges that we may*

			face in implementing VAMs is a critical shortage of both financial and human resources…." (PM2) *"One challenge that we may face in implementing VAMs is a lack of human resources with expertise in statistical techniques…." (PM3)* *"The new evaluation method would impose an extra workload on schools…" (HT1)* *"We recognize the importance of using data to inform school improvement, but we need staff who are able to analyse and interpret data accurately and communicate the results in a way that is accessible and meaningful to all staff." (HT1)* *"No single evaluation information can be a perfect indicator of school performance; therefore, the value-added evaluation results can be considered a supplementary indicator of school performance. "(PM2)* *"I think that VAMs results should be the primary measure of school performance, as they are perceived to be a more robust and reliable school evaluation method. Therefore, the VAMs results can provide a more comprehensive and accurate assessment of the school's impact on student learning and progress over time."* (PM3)