Assessment of Primary Chinese EFL Learners' Oral and Aural Language Skills and Their Relations to the National English Curriculum Standards Targets



Xi Lian

Jesus College

Department of Education

University of Oxford

Thesis submitted to the University of Oxford in fulfilment of the requirements for the Degree of Doctor of Philosophy

Michaelmas 2018

Abstract

In recent years, the governments of many non-native English-speaking countries are incorporating English as a foreign language (EFL) programmes into primary school curricula at increasingly lower grade levels. However, despite an increasing number of young EFL learners around the world, relatively little contextualised research has examined the assessment practices and the learning outcomes of young EFL learners at the primary level, especially in East Asian countries. This study aimed to investigate the assessment of the oral and aural language abilities of Chinese primary EFL learners from two key grade levels (i.e., Grade 4 and Grade 6) in relation to the set targets articulated by the National English Curriculum Standards (NECS). More specifically, this research examined how teachers actually assessed their students' oral and aural abilities in classrooms; which assessment activities were used in the classroom and how these activities related to the NECS targets; and what outcomes students could actually accomplish regarding speaking and listening upon finishing Grade 4 and Grade 6. To provide an in-depth, contextualised and comprehensive picture of both the assessment process and its outcomes in the Chinese primary school instructional setting, a mixed-methods research design was employed, combining quantitative analysis of teacher and student questionnaires and student individual assessments with qualitative analysis of classroom observations, and teacher and student focus-group interviews. Four EFL teachers, 106 Grade 4 students and 116 Grade 6 students from two local Chinese state primary schools participated in this study.

The main findings are as follows. First, the EFL teachers tended to assess students' listening comprehension and oral production skills at the level of both vocabularies and sentences; however, students' oral and aural skills at the discourse level were relatively less frequently assessed. In addition, the assessment activities that were employed were organised as teacher-centred, and they were textbook-dominated. Second, two main sets of factors, subdivided into external and internal factors, were found to influence teachers' interpretation and implementation of the curriculum targets. The external factors included the provision of adequate resources and support; contextual constraints (i.e., large class sizes with mixed-level students, limited instructional hours) and washback effects of the junior entrance exams. The internal factors mainly concerned teachers' English proficiency and their professional pedagogical knowledge and skills. Third, in terms of students' actual performance on the two external individual assessments in alignment with the NECS target levels, fewer than half of the students in Grade 4 and Grade 6 were judged by a panel of local expert teachers to have achieved the NECS listening targets at Level 1 and Level 2, respectively. With regard to speaking skills, fewer than half of the Grade 4 students were classified as qualified candidates at Level 1, but more than half of the Grade 6 students were judged as qualified candidates at Level 2. Therefore, the study's overall findings indicated that the NECS targets were underachieved in the two sampled state primary schools; furthermore, these findings presented a number of implications for pedagogy, resources and support, and assessment, as well as directions for future research.

Acknowledgement

The completion of this thesis would not have been possible without the support and encouragement from many people.

First of all, I would like to express my heartfelt gratitude to my two supervisors Professor Victoria Murphy and Dr Maria Evangelou for their intellectual inspiration, continuous support and insightful guidance. Professor Victoria Murphy not only developed my critical thinking and reasoning but also showed me what a real independent and successful scholar should be. Dr Maria Evangelou always gave me timely and encouraging comments. I feel very fortunate to have such inspiring and also caring supervision over the course of this DPhil journey at Oxford.

I would also like to thank all those academics who guided me through every milestone of my research, including Professor Pam Sammons, Dr Robert Woore, Dr Heath Rose, Dr Joshua McGrane and Professor Janet Enever, for their invaluable suggestions and encouraging comments. Additionally, I am especially grateful to all members of the REAL group and the colleagues in the Department of Education and Jesus College, who have supported me in every possible way and have made my doctoral life a meaningful and enjoyable one.

My gratitude also goes to all the teachers and students who participated in my study and to the school principals and staff members who trusted me and allowed me to enter the classrooms. Without their cooperation and understandings, the study would not have been possible.

Finally, on a personal note, I would like to thank all my wonderful friends and beloved family members, for their unconditional support, trust and love; Especially to my dearest father and mother, I am truly blessed to be your daughter. This thesis is dedicated to you.

Table of contents

Abstract **1**

List of tables **8**

List of figures **10**

Chapter 1 Introduction **11**

1.1 Introduction 11

1.2 Research background 11

1.2.1 Why focus on EFL learning in the primary school context? 11

1.2.2 Why focus on assessment? 12

1.2.3 Why focus on English aural and oral language skills? 13

1.3 Research aims 14

1.4 Research questions 14

1.5 Overview of the thesis 16

Chapter 2 Literature review **18**

2.1 Introduction 18

2.2 Young language learners 18

2.2.1 Young language learners in the EFL setting 18

2.2.2 The special characteristics of young language learners 18

2.3 Early EFL programmes 21

2.3.1 The younger the better? 21

2.3.2 Early EFL programme curriculum policies and their implementation in East Asian countries 23

2.3.3 The challenging factors of curriculum implementation 28

2.4 Assessment of young EFL learners 31

2.4.1 Constructs of language ability in relation to assessments in listening and speaking 31

2.4.2 Assessing the process: classroom-based assessment 33

2.4.3 Assessing the outcomes 38
2.5 Summary of chapter two 42
Chapter 3 Research methodology 44
3.1 Introduction 44
3.2 Research rationale and mix-methods research design 44
3.3 Sampling 48
3.3.1 Rationale for sample selection 48
3.3.2 Procedures of sample selection 48
3.3.3 Participants 49
3.4 Research instruments 52
3.4.1 Questionnaires 52
3.4.2 Interviews 57
3.4.3 Classroom observations 59
3.4.4 Assessments 63
3.5 Data collection procedures 77
3.5.1 Data collection procedure for questionnaires 77
3.5.2 Data collection procedure for classroom observation 77
3.5.3 Data collection procedure for interviews 78
3.5.4 Data collection procedure for individual assessments 78
3.6 Data analysis 79
3.6.1 Data analysis of questionnaire 79
3.6.2 Data analysis of interview 80
3.6.3 Data analysis of classroom observation 81
3.6.4 Data analysis of assessments 81
3.7 Ethical considerations 82
3.8 Summary of chapter three 83

Chapter 4 Data Analyses and results: classroom context 84

4.1 Introduction 84

4.2 Teacher's commonly used classroom-based assessment activities in assessing listening and speaking skills 84

4.2.1 Results of the teacher questionnaire 84

4.2.2 Results of classroom observations 86

4.3 Teacher classroom-based assessment practices 97

4.3.1 Results of teacher questionnaire 97

4.3.2 Results of classroom observations 99

4.4 Summary of chapter four 104

Chapter 5 Data analyses and results: teachers' perspectives 106

5.1 Introduction 106

5.2 Teachers' general perceptions of the NECS 106

5.2.1 Teachers' views on NECS implementation: results from the teacher questionnaire 106

5.2.2 Teachers' views on NECS implementation: results from the teacher interviews 109

5.2.3 Summary 119

5.3 Teachers' attitudes towards the NECS targets in listening and speaking 120

5.3.1 Results from the teacher questionnaire 121

5.3.2 Results from the teacher interviews 124

5.3.3 Summary of chapter five 125

Chapter 6 Data Analyses and results: student outcomes 126

6.1 Introduction 126

6.2 Young EFL learners' English oral and aural language proficiency at primary level 126

6.2.1 EFL students' performance results on two distinct assessments 127

6.2.2 Linking individual assessment scores with the NECS 130

6.3 Students attitudes towards EFL learning and assessment 141

6.3.1 Students' attitudes towards EFL learning 141

6.3.2 Student preferences for classroom-based aural and oral assessment activities 145
6.4 Summary of chapter six 150
Chapter 7 Discussion and conclusions 151
7.1 Introduction 151
7.2 Synthesis of the findings 151
7.2.1 The findings for RQ1: Inside the *Black Box* 151
7.2.2 The findings for RQ2: Factors related to classroom-based assessment implementation 160
7.2.3 The findings for RQ3 and RQ4: students learning outcomes by Grade 4 and Grade 6 163
7.3 The significance and implications of the study 169
7.3.1 Implications for pedagogy 169
7.3.2 Implications for resources and for teachers' professional training 171
7.3.3 Implications for assessors 174
7.4 Limitations and suggestions for the future study 175
7.5 Final conclusions 177
References 178
Appendices 193
Appendix 1 Level descriptors of language skills in the NECS 193
Appendix 2 Teacher questionnaire 195
Appendix 3 Student questionnaire 200
Appendix 4 Teacher interview schedule 203
Appendix 5 Student interview schedule 205
Appendix 6 The adapted COLT scheme 206
Appendix 7 British Ability Scale II: Verbal Comprehension record booklet 207
Appendix 8 Storytelling task 209
Appendix 9 Index of Narrative Complexity Story Coding Form 212
Appendix 10 Invitation letter to school principal 219

Appendix 11 Information sheet for teachers 221

Appendix 12 Information sheet for parents/guardians 223

Appendix 13 Consent form for teachers.....226

Appendix 14 Consent form for parents/guardians.....227

List of tables

Table 2. 1 Overview of the EFL curriculum policies and its provisions at the primary school level in South Korea and Japan 24

Table 2. 2 The new NECS goals 27

Table 3. 1 Summary of research questions, participants and instruments……...46

Table 3. 2 Demographic information of teacher participants 50

Table 3. 3 Background information of teacher participants 52

Table 3. 4 Demographic information on student participants 52

Table 3. 5 Structure and themes of the teacher questionnaire 56

Table 3. 6 NECS targets in Level 1, contributing knowledge and skills, and research instruments 65

Table 3. 7 NECS targets in Level 2, contributing knowledge and skills, and research instruments 66

Table 3. 8 Descriptive statistics for Grade 4 students in pilot study 69

Table 3. 9 Descriptive statistics for Grade 6 students in pilot study 70

Table 3. 10 Demographic information on panelists 73

Table 4. 1 The frequency-ranked list of classroom-based assessment activities used by teachers in assessing listening and speaking skills……...85

Table 4. 2 The frequency-ranked list of two Grade 4 teachers' classroom-based assessment activities 88

Table 4. 3 Time and percentage of time spent on different categories by two Grade 4 teachers.....89

Table 4. 4 The frequency-ranked list of two Grade 6 teachers' classroom-based assessment activities 93

Table 4. 5 Time and percentage of time spent on different categories by two Grade 6 teachers.....94

Table 4. 6 Agent and focus of assessment practices.....98

Table 4. 7 Time and percentage of time spent on different categories by two Grade 4 teachers.....99

Table 4. 8 Time and percentage of time spent on different categories by two Grade 6 teachers.....103

Table 5. 1 Results of factor analysis for the teachers' general perceptions of curriculum implementation in aural and oral teaching……..108

Table 5. 2 Descriptive statistics and internal consistency for the factors of teachers' general perceptions of curriculum implementation.....109

Table 5. 3 Summary of factors influencing teachers' perceptions of curriculum implementation from the teacher interview and questionnaire 120

Table 5. 4 Teachers' perceptions of NECS targets at Level 1 122

Table 5. 5 Teachers' perception of NECS targets at Level 2.....123

Table 6. 1 Descriptive statistics for Grade 4 students……127

Table 6. 2 Descriptive statistics for Grade 6 students 127

Table 6. 3 The mean ranked list of Narrative elements scores.....129

Table 6. 4 Independent samples t-test of students INC scores of narrative elements by grade 130

Table 6. 5 General descriptors and panelists' descriptors of NECS Level 1 listening skills 131

Table 6. 6 General descriptors and panelists' descriptors of NECS Level 2 listening skills 132

Table 6. 7 Panelists' cut-off score judgements on BAS-II Verbal Comprehension, Level 1.....133

Table 6. 8 Panelists' cut-off score judgements on BAS-II Verbal Comprehension, Level 2.....134

Table 6. 9 Distribution and t-test of Grade 4 students by listening skill level 135

Table 6. 10 Distribution and t-test of Grade 6 students by listening skill level 135

Table 6. 11 General descriptors and panelists' descriptors of NECS Level 1 speaking skills.....136

Table 6. 12 General descriptors and panelists' descriptors of NECS Level 2 speaking skills.....136

Table 6. 13 Confusion/error matrix (Kosuke Imai, 2017).....139

Table 6. 14 Decision table for potential cut-off scores for the frog storytelling test at Level 1.....139

Table 6. 15 Decision table for potential cut-off scores for the frog storytelling test at Level 2.....140

Table 6. 16 Distribution of Grade 4 students by speaking skill level.....140

Table 6. 17 Distribution of Grade 6 students by speaking skill level.....141

Table 6. 18 Results of factor analysis for the students' affective attitudes towards EFL learning.....143

Table 6. 19 Descriptive statistics and internal consistency for the students' attitudes factor scales ...144

Table 6. 20 Independent samples t-test of students mean responses to affective factor scales by grade 144

Table 6. 21 The mean ranked list of classroom-based aural and oral assessment activities by Grade 4 and Grade 6 students.....145

Table 7. 1 A set of questions for classroom teachers to follow during the process of assessment……..171

List of figures

Figure 2. 1 Wiliam and Thompson’s framework of formative assessment in class.....35

Figure 3. 1 Convergent parallel model mixed-methods design (Creswell & Clark, 2007)……45

Figure 6. 1 Frequency distributions of Grade 4 students' frog storytelling test scores in 2 categories...137

Figure 6. 2 Frequency distributions of Grade 6 students' frog storytelling test scores in 2 categories...138

Chapter 1 Introduction

1.1 Introduction

This chapter sets up the research background for the current study. Then it specifies the research aims and presents the research questions. After that, it gives an overview of the whole thesis.

1.2 Research background

This section outlines background information related to the current study and research topic, namely, the assessment of oral and aural language skills among primary school-aged English-language learners in China, and the relationship of these assessments to targets set by the National English Curriculum Standards (NECS).

1.2.1 Why focus on EFL learning in the primary school context?

Primary school children across the world start learning foreign languages, especially English, as part of their schools' curriculums. The governments of some non-native English-speaking countries (e.g., China and Korea) are incorporating English as a foreign language (EFL) programmes into primary school curricula and are lowering the starting age for these programmes. According to a recent international survey of policies and practices for teaching English at the primary level, nearly more than one-half countries surveyed (34 out of 64 countries) had introduced English instruction in the Grade 1 curriculum in the 12 years since the administration of a prior survey on this subject (Rixon, 2013). Driven by the assumption that earlier is better for English instruction, policymakers in these countries believe that starting at an earlier age will enhance their citizens' English language skills and increase their social mobility, bringing national economic prosperity in a globalised economy (Enever, Moon, & Raman, 2009; Hayes, 2014; Nikolov & Mihaljević Djigunović, 2011). Although some studies have indicated that children benefit cognitively from an early start, in these studies second language acquisition took place in a natural setting rather than a school (Genesee, Lindholm-Leary, Saunders, & Christian, 2006; Skehan, 1998). However, learning EFL in a school is quite different. Simply lowering the starting age does not guarantee successful learning outcomes. There are other, arguably, more important contextual factors that interact with the age of acquisition, including the quality, amount and continuity of input, the available instruction time, EFL teachers' proficiency and the support EFL teachers receive from their schools (Butler, 2009a; Dekeyser, Alfi-Shabtay, & Ravid, 2010; Murphy, 2014; Nikolov & Mihaljević Djigunović, 2006; Nunan, 2003). At present, there has been little research into actual learning outcomes under these language policies. Specifically, contextualised research that focuses on what students can actually do in terms of their English proficiency, as well as on students' language learning experience in primary school, is emerging but remains scarce (Enever, 2011;

Khabbazbashi, Khalifa, Robinson, Ellis, & Mifsud, 2016). In China, English was first introduced as a compulsory subject at the primary level in 2001; however, after 17 years curriculum implementation, teachers' classroom practices and students' learning outcomes, especially in relation to curriculum targets, remain unresearched. Given this context, this study aims to take a snapshot of Chinese students' learning outcomes in state primary schools.

1.2.2 Why focus on assessment?

As discussed in the previous section, in order to get a general picture of what Chinese primary EFL students can do in relation to the national curriculum, assessment is key. Here, the assessment in the research title refers to two aspects of assessment: the assessment process and assessment outcomes.

First, the issues regarding the *assessment process*. Previous studies have noted that intended curriculum goals often do not correspond to actual classroom practice (Baldauf, Kaplan, & Bryant, 2011; Nunan, 2003). In order to have a comprehensive picture of implementation for early EFL programmes and to better understand primary EFL students' learning outcomes, it is essential to go inside the *'black box*' (i.e., the classroom) to find out how the actual classroom assessment process works (Black & Wiliam, 1998b). Recent studies have gradually broadened the definitions and functions of assessment beyond the traditional model of externally set tests designed with selection or accountability in mind (Black & Wiliam, 2009; Davison & Leung, 2009; Hill & McNamara, 2011; Rea-Dickins & Gardner, 2000). Assessment types include assessments for learning, formative assessments, classroom-based assessments and dynamic assessments, which are treated as an integral part of the teaching process that facilitates learning and results in learning gains. Researchers have explored the process, elements and strategies of effective assessment (Leung, 2007; Rea-Dickins & Rixon, 1999; Wiliam & Thompson, 2007). However, there is still a lack of empirical research on the assessment of primary EFL learners in the classroom context. In particular, there is a lack of empirical data regarding which types of classroom-based assessment activities are used, how teachers implement these activities and how both the types of activities and their implementation relate to intended curriculum goals and objectives.

Second, there is the question of *assessment outcomes*. In addition to exploring the assessment process in class, it would also be important to know whether the curriculum goals and objectives have been achieved by the students. In an alignment manner, if the curriculum goals are designed to promote language use, then the students should have opportunities to participate in activities that involve meaningful language use. Assessment should also focus on the area of language use ability (McKay, 2006). In many East Asian countries (e.g., Japan, Korea, China), the main goal of early EFL programmes are to enable young EFL learners to master language use ability especially, to be able to use the target language for communicative purposes in the real-life context. Therefore, the term *assessment outcomes* does not refer to the scores measured by the traditional paper-and-pencil tests that

focus on spelling, vocabulary, and grammar. The term *assessment outcomes* refers mainly to language use ability as defined in the curriculum. However, in the current East Asian context, which is dominated by traditional-paper-and-pencil tests and which seems to lack standardised tests designed in alignment with curriculum goals, little is known about students' actual language use learning outcomes (Butler, 2015). In addition, these early EFL programmes also encourage young EFL leaners to develop favourable attitudes towards foreign language learning (Nikolov & Mihaljević Djigunović, 2006). Doing to is important because students' attitudes towards learning a foreign language might influence the students' behaviour with regards to both how much effort they are willing to invest and how long they are willing to persist in the face of difficulties (Csizer & Dörnyei, 2005; Kormos & Csizér, 2008); importantly, these attitudes include students' beliefs about their capacity to learn as well as the students' perceptions of their learning experience and the extent of these perceptions. Therefore, assessment might need to address learning outcomes at two levels: linguistic and non-linguistic. Linguistic outcomes refer to young EFL learners' language achievement at a certain stage, and non-linguistic outcomes refer to young EFL learners' affective attitudes towards their language-learning experience. These two outcomes cannot be teased apart when assessing young EFL learners' language achievements (Enever, 2011; Khabbazbashi et al., 2016).

1.2.3 Why focus on English aural and oral language skills?

The reasons for focusing on English aural and oral language skills are as follows. First, at the early stages of foreign language learning, especially for young EFL learners, oracy (i.e., listening and speaking) is developed first, followed by literacy (i.e., reading and writing). This progression allows children to engage in classroom activities and to interact and receive feedback from teachers and peers; it also lays the foundation for later literacy development (McKay, 2006). Second, communicative competence has been widely accepted as one of the main curriculum goals in many early EFL programmes, as it not only focuses on linguistic aspects but also emphasises the ability to use English for communicative purposes (Bachman & Palmer, 1996). In China, the national curriculum has placed equal emphasis on oracy and literacy skills, encouraging the use of English in real-life situations (Chinese Ministry of Education, 2011).

However, the assessment of young EFL learners' oral language is still little researched due to the challenges that remain: (1) it is difficult to find or design age-appropriate instruments, and it is challenging to capture children's incremental oral progress at the initial stage of EFL learning (Enever, 2011; McKay, 2006); (2) the educational system does not align with the equal emphasis on both oracy and literacy that stated in the national curriculum. In China, the high-stake examination-oriented educational system remained unchanged and the examinations usually in the traditional paper-and-pencil format and do not assess oral language. Therefore, this would exert the exam washback effect on

daily classroom teaching in which teachers are under the exam pressure to teach for examination rather than for communication (Li & Baldauf, 2011).

1.3 Research aims

The overarching objective of this study is to investigate the assessment of the oral and aural language ability of Chinese young EFL learners at the primary school level in relation to the set targets articulated by the national curriculum. More specifically, this study explores the implementation of the mandatory National English Curriculum Standards (NECS) in China by looking at teachers' practices and their interpretations of the intended curriculum in their classrooms, as well as at students' actual learning outcomes. This research is carried out at two key grade levels (i.e., Grade 4 and Grade 6). It intends to gain an in-depth, contextualised and comprehensive picture of the assessment of young EFL learners in the Chinese primary school instructional setting.

Specifically, this research is aimed at the following:

1) investigating how teachers actually teach and assess their students' oral and aural ability in class, which activities the teachers use and how these activities relate to the NECS targets;

2) finding out the two stakeholders' (i.e., teachers' and students') perceptions and attitudes towards EFL teaching/learning and assessment;

3) examining what students can actually do in terms of listening and speaking, as will be documented via two external assessments and how a group of local teachers interpreted their students' performance in relation to the NECS targets via the standard setting methods.

1.4 Research questions

The present study is guided by the following key research questions (RQs) organised around the stakeholders in the classroom: teachers and students. The study mainly touches on the following three aspects: the classroom context; teachers' perspectives and students' outcomes.

This part attempts to describe what is going on in the classroom, as the first layer of investigation, by providing factorial data at two levels of the target context. First, at the activity level, the study aims to explore which types of classroom-based aural and oral assessment activities are being used and which features are associated with these frequently used assessment activities. Second, at the practice level, it aims to investigate how teachers implemented these activities in the classroom.

Part I: Classroom context

RQ1. How do Grade 4 and Grade 6 EFL teachers in state primary schools conduct their assessments in listening and speaking in the classroom context in relation to the NECS targets?

RQ1.1 Which classroom-based assessment activities do EFL teachers use for students in Grade 4 and Grade 6?

RQ1.2 How are these assessment activities implemented in the classroom?

Part II: Teachers' perspectives

This part focuses on teachers' perspectives towards curriculum implementation, specifically with regard to aural and oral teaching. The study aims to explore the potential factors that facilitate or inhibit teachers' perceptions and practices in relation to the intended curriculum.

RQ2. How is the intended NECS being interpreted by local EFL teachers?

RQ2.1 What are teachers' general perceptions of the curriculum implementation in aural and oral teaching?

RQ2.2 To what extent do teachers believe that their students will reach the curriculum targets in listening and speaking skills by the end of Grade 4 and Grade 6?

Part III: Students' outcomes

This part intends to investigate students' actual learning outcomes in relation to set targets on the NECS at two key grade levels (i.e., Grade 4 and Grade 6). First, students' aural and oral language skills are assessed via two individual assessments. Then, the students' test results are matched with the NECS targets via interpretations made by a group of local teachers. Next, the students' affective attitudes, as another layer of learning outcomes, are investigated. Specifically, the students' general affective attitudes towards EFL learning at school, as well as the students' opinions of frequently used classroom-based aural and oral assessment activities, are also explored.

RQ3. According to the targets set by the NECS, what can young EFL learners actually do in speaking and listening upon finishing Grade 4 and Grade 6 respectively?

RQ3.1 What are Grade 4 and Grade 6 EFL students' performance results on the two individual assessments, the BAS-II verbal comprehension and the Frog storytelling task?

RQ3.2 How do a group of teachers interpret the Grade 4 and Grade 6 students' results on the two individual assessments in relation to the levels (i.e. Level 1 and Level 2) in the NECS?

RQ4. What are the attitudes of young EFL learners towards the learning, teaching and assessment of EFL in Chinese primary state schools?

RQ4.1 What are the affective attitudes of young EFL learners at two key grade levels (i.e., Grade 4 and Grade 6) towards EFL learning?

RQ4.2 What are young EFL learners' preferences for classroom-based aural and oral assessment activities at two key grade levels (i.e., Grade 4 and Grade 6)?

1.5 Overview of the thesis

The thesis is organised into seven chapters that collectively describe the research conducted. Below is a brief outline of the chapters in the thesis.

Chapter 1 introduces the background of the research, addressing questions related to the research title. Then the chapter discusses the research aims and research questions. It ends with an overview of the thesis as a whole.

Chapter 2 critically reviews the relevant literature, explaining key definitions and concepts and reviewing prior studies based on the following key issues: special characteristics of young EFL learners; an overview of primary EFL learning programmes in formal instructional settings; factors related to curriculum implementation; assessment of oral and aural skills; and classroom-based assessment and externally used assessments. The chapter concludes by outlining the theoretical and empirical grounds of the current study.

Chapter 3 presents a detailed description of the research methods used in this study. It starts with the research rationale and the mixed-methods research design, followed by the sampling frame. After that, the reasons for adopting each research instrument used for the main data collection are presented in together with a discussion of the adaptation process for the instruments. An explanation of the research procedure and data analysis is followed. In addition, ethical considerations are addressed.

Chapters 4, 5 and 6 report the research findings. Chapter 4 addresses the first main research question by providing an overview of the classroom data in Grade 4 and Grade 6 in a sample of 2 state primary schools. Chapter 5 addresses the second main research question on teachers' perspectives. It presents findings on teachers' perceptions of the NECS targets and the teachers' actual implementation of the

curriculum in their classrooms. Chapter 6 focuses on reporting students' aural and oral learning outcomes in relation to curriculum targets.

Chapter 7 brings together findings from the previous 3 chapters. It summarises the issues emerging from the findings, which are then discussed in relation to the focus of the research and the existing literature. The potential significance and implications of the current research are also stated. Limitations of the current study and issues for future research are also discussed in the final chapter.

Chapter 2 Literature review

2.1 Introduction

This chapter starts by describing the special characteristics of young language learners and then moves on to an overview of early EFL programmes in formal instructional settings, particularly in an East Asian context. Next, the chapter reviews the analysis of a number of factors identified in prior studies that influence curriculum implementation and, potentially, influence children's learning outcomes. This chapter then discusses the assessment of young learners, which is seen as a bridge between curriculum policies and students' learning outcomes. This assessment is reviewed and discussed at two levels: assessing the process and assessing the outcomes. Key concepts in assessment, as well as related studies in classroom-based assessment, are reviewed. In addition, externally used assessment instruments are discussed.

2.2 Young language learners

2.2.1 Young language learners in the EFL setting

Young EFL language learners are defined here as children, ages 6 through 12, who are learning English as a foreign language in a primary school setting; in the context of East Asia, children normally start primary education at the age of 6 or 7 (Grade 1) and finish at the age of 11 or 12 (Grade 6) (Butler, 2015; McKay, 2006). Young EFL learners take English as a mandatory subject, and their contact with the language is generally limited to the classroom environment. This formal instructional context differentiates these learners from those learning English as a second or additional language (i.e., ESL or EAL), who are usually living in a community in which English is the primary language and is used not only in educational settings but also in the community for the purpose of communication (McKay, 2006).

2.2.2 The special characteristics of young language learners

As McKay (2006) noted,

> Children bring to their language learning their own personalities, likes and dislikes and interests, their own individual cognitive styles and capabilities and their own strengths and weaknesses. (p. 5)

Such individuality during children's language learning differentiates young learners from their adult counterparts. Therefore, it is important to review what is known about young learners' language learning in order to make meaningful connections with effective assessment, specifically in instructional settings. According to McKay (2006), young learners' characteristics are categorized into

three main categories: (1) cognitive, social and emotional, and physical growth; (2) literacy; and (3) vulnerability. In the current study, since the primary focus is on the assessment of primary language learners' oral and aural abilities, consideration of literacy in relation to reading and writing is omitted. Therefore, only two main categories are reviewed: (1) cognitive and physical characteristics and (2) affective characteristics (i.e., social, emotional and vulnerability).

2.2.2.1 Cognitive and physical characteristics

Unlike adult learners, young language learners experience significant cognitive and physical growth (Bialystok, 2001; McKay, 2006). During lower grade levels, children learn and increase their understanding through direct experiences, such as through actions, real objects or visual aids. In addition, children still find it difficult to analyse language and rely on their memorisation to organise and recall information (Puckett & Black, 2000). As children move on to the upper grade levels, though they still rely on hands-on experience, their thoughts and ideas display less egocentrism, becoming less subjective and more objective. The children also tend to see a task as a whole, instead of focusing on only one aspect while ignoring others (Berk, 2000; McKay, 2006). As young learners age out of primary education, around the age of 11 or 12, they begin to understand cause and effect, temporal events, and abstract concepts. They also begin to develop the ability to form hypotheses and deductive reasoning (Puckett & Black, 2000).

Such progressive cognitive differences between younger and older language learners were also found among foreign or second language learners. A dual-mode system theory in language use and learning was proposed by Skehan (1998). In this model, two systems were found to coexist during language processing: a formulaic system (also known as an exemplar-based system) and a rule-based system. Young language learners in the lower primary grades were found to at the beginning school stage rely mainly on the formulaic system, that is, on understanding and using ready-made language chunks, such as formulaic expressions and, sometimes, well-formed sentences. These language patterns were usually grammatically unanalysed, meaning that the young learners understood or used these linguistic strings as a whole; the young learners could not analyse these strings into parts or understand the underlying rules involved. For example, while these young language learners could understand commonly used teaching instructions (e.g., "Let's listen to the radio and answer the questions."; "Please put your hands up if you have finished.") during class and act correspond accordingly, they could not recognise grammatical rules. These learners' understanding of language chunks was also are usually "context-bound," meaning that these young learners could easily understand a language chunk in a specific context but could not apply that same chunk to other contexts (Skehan, 1998, p. 89). Thus, although Skehan (1998) indicated that adult learners or first language speakers usually use the rule-based system (i.e., to employ grammatical rules) at the same time to construct sentences and discourses, this tended

to occur in a well-developed condition (e.g., a first language speaker's system). Young language learners, in contrast, are still developing their metalinguistic ability, and their language learning is not yet mature, therefore, at the initial stage, they hardly rely on the rule-based system. For young learners in the upper primary grade levels, these learners may begin, as they accumulate language chunks, to consciously gain language rules from explicit language study. However, whether the rule-based system could be said to be effectively activated and to work together with the formulaic system depends on the actual learning context (McKay, 2006).

Regarding young learners' physical characteristics, during the lower grade levels, students love to engage in activities that involve physical responses and actions; however, these young learners' attention span is usually short, and they easily get tired (McKay, 2006). At the upper grade levels, these students have better hand-eye coordination and a longer attention span.

Therefore, knowledge of young learners' cognitive and physical characteristics is important for the effective assessment of young learners, especially for selecting and constructing appropriate assessment tasks and creating optimal assessment conditions. As suggested by McKay (2006), for younger language learners, assessment tasks should be familiar to the students. The assessment tasks should probably also contain repetitive elements so younger learners can draw upon the formulaic system to facilitate their participation. Such tasks should also involve physical responses to help motivate the students to complete the tasks. In addition, the focus of the assessment tasks should be on meaning, rather than form. In contrast, for older language learners, more rule-based assessment activities that require longer concentration spans could be used to assess learners' use of grammar and vocabulary; however, these activities still need to be embedded into a meaningful context which the children are familiar with and can handle.

2.2.2.2 Affective characteristics

Affective characteristics primarily address the emotional aspects of young language learners as well as the social aspects of these learners' environments. Unlike learning a language at home, primary students are learning a language in a context (i.e., a classroom) in which they are in frequently contact with teachers and peers. At the initial stage of primary education, these students both develop their social awareness and learn how to cooperate and to share thoughts and ideas with peers. In addition, these students display high motivation (Szpotowicz, Mihaljević Djigunović, & Enever, 2009), overt confidence (McKay, 2006) and low levels of anxiety (Nikolov & Mihaljević Djigunović, 2011). As they grow older, they continue to develop their teamwork ability. At the same time, the students' motivation and confidence decline (Lopriore & Mihaljević Djigunović, 2009). The students become sensitive to comments made by their teachers and peers, and they become vulnerable to criticism and failure (McKay, 2006). These findings have implications for assessment. First, assessment activities

should be conducted in small groups, and students should have opportunities to interact with their peers. Second, the selection of assessment tasks needs to ensure that the children are engaged in a successful language learning environment in which their positive language learning motivation and interests are evoked (Nikolov & Mihaljević Djigunović, 2011). For example, teachers or assessors need to build a certain amount of flexibility into assessment tasks so that children at different proficiency levels can experience some degree of success. Third, assessment feedback from teacher and peers should be immediate and positive, providing a safe and supportive learning environment (McKay, 2006). The implications of the preceding discussion will be considered for the present study when selecting and administering assessment instruments for young language learners.

2.3 Early EFL programmes

Having reviewed young language learners' special characteristics, this section moves on to discuss the instructional contexts in which young learners study English as their first foreign language. This section consists of three parts. The first part discusses the correlation between age and learning outcomes for young language learners. The second part reviews current early EFL programme policies and provisions in East Asian countries. The third part reviews potential factors that influence curriculum implementation and students' actual learning outcomes.

2.3.1 The younger the better?

Before reviewing current EFL programmes and the provisions they offer, this section discusses the age of acquisition effect. The prevailing assumption 'the younger the better', meaning that the younger the learner is, the better the learning outcomes will be, has persuaded policymakers both to introduce EFL programmes into the curriculum at the initial phase of formal schooling and to lower the starting age for such programmes. The theory behind this assumption comes from the Critical Period Hypothesis, or CPH, which claims that children have an innate advantage (over older adolescents or adults) for learning a language. According to the CPH, children's brain plasticity is believed to facilitate language learning; as a result, children find it easier to learn a language than older adolescents or adult learners do. The CPH claims that this advantage begins around age 2 and lasts until age 12 (i.e., puberty) (Lenneberg, 1967), thus identifying a critical period after which any language learning would be less successful. However, the original claims made under CPH now seem problematic since recent evidence has indicated that there are no exact cut-off points for such period (Hakuta, Bialystok, & Wiley, 2003; MacWhinney, 2005). Recent evidence from the neurological studies using the brain imaging technologies (e.g. functional magnetic resonance imaging or fMRI) though claimed that the particular brain region (e.g. grey matter) of children who were being exposed to and acquiring additional languages at an early age was expanded in comparison with their monolingual counterparts. The studies further indicated such benefits derived from young children's brain plasticity might gradually decline

as the age increases (Mechelli et al., 2004; White, Hutka, Williams, & Moreno, 2013).

Some research have corroborated these age-of-acquisition effects, arguing that children do enjoy some advantages in learning a second language based on their implicit learning mechanisms, in contrast to adults' explicit mechanisms (Muñoz, 2006). This research has also found that, in the long run, children can achieve native levels of proficiency in pronunciation (Flege, Munro, & MacKay, 1995; Long, 2005) as well as grammar (DeKeyser, 2000). However, there are problems applying this research to the EFL context because these research studies focused on children from migrant families living in 'natural language environments', meaning that the second language was used not only in the classroom but also in the community (Dekeyser, 2012). When the same research was applied to the EFL formal schooling setting, the benefits of an early start could not be guaranteed as they were in the naturalistic context. For example, the Barcelona Age Factor Project (BAF), a large-scale and longitudinal study, aimed to compare whether older learners who started EFL instruction at age 11 could surpass their younger counterparts who started at age 8 in the EFL setting (Muñoz, 2006). The data was collected after each group of learners had amassed around 700 hours of EFL instruction at school. A number of measures, including tests of four macro-language skills (i.e., listening, speaking, reading and writing), were used. The findings revealed that at the beginning stage (i.e., at the start of instruction), older learners outperformed their younger counterparts on all measures. After nine years of EFL instruction, however, the learners who began study at a younger age gradually caught up on all measures, especially on the test in implicit learning. Given the limited input from the EFL instructional setting, it might be concluded that the long-term effects of an early start might not be fully facilitated. The results of a later longitudinal study, which further investigated the influence of starting age and input on EFL learners' oral performance (Muñoz, 2014), confirmed the previous BAF findings that there was a lack of association between starting age and length of instruction with proficiency measures. Instead, the quality of input and the availability of contacting with native speakers were showed to be determinant predictors for learners' oral performance. Similarly, other studies have also confirmed that, given the opportunity for extensive exposure to a foreign language, early starters could outperform older ones (Djigunović, Nikolov, & Ottó, 2008; Nikolov & Curtan, 2000).

Therefore, it may be that the age-of-acquisition effect is not context-independent. Rather, there are other contextual variables, like the quality and quantity of language input; the types and continuity of educational language programmes and the instructed learning environment. These variables could not be teased apart when considering the foreign language learning outcomes (Hakuta et al., 2003; Murphy, 2014; Nikolov & Mihaljević Djigunović, 2006; Singleton, 2014). Moreover, as suggested by Murphy (2014), given the limited-input environment, it would be reasonable and meaningful to understand the context of young language learners in terms of "the realistic foreign language learning outcomes as well as the nature of learning in these contexts" (Murphy, 2014, p. 24). Thus, in the next section, in order to

get a relatively comprehensive picture of the broad context of young EFL learners in East Asia, it is worth briefly reviewing which policies are currently implemented, which provisions are available, and, most importantly, which assessments or learning outcomes are employed.

2.3.2 Early EFL programme curriculum policies and their implementation in East Asian countries

On a global scale, the growing need for English and its underlying advantages for the global economy, as well as for individual intercultural communication competence, have motivated many countries to include English as a compulsory part of the primary curriculum. There is also an evident pattern of policies rapidly lowering the starting age of instruction (Enever, 2018). East Asian countries are no exception. This section provides an overview of current curriculum policies, along with their provisions and the assessments used in some East Asian countries. The section first reviews the policies and provisions of China's neighbouring countries, South Korea and Japan, and then moves on to review the policies and provisions in China. Problems with the implications of the EFL curriculum policies in these countries are also discussed.

2.3.2.1 Early EFL programme curriculum policies and their provisions in South Korea and Japan

South Korea and Japan were chosen for the present study for the following reasons. First, both countries belong to the Expanding Circle, in which English is taught and learned as a foreign language, and in which people have rare opportunities to use English for communicative purposes within their communities (Kachru, 1992). Second, both South Korea and Japan are neighbouring countries to China, and the two countries share similar socio-historical contexts in terms of English language learning and teaching. Third, given the limited research in formal EFL learning at the primary level in East Asia, previous studies in South Korea and Japan can provide information to better understand the situation in China. The following key issues are considered when summarizing the policies and provisions of early EFL programmes at the primary school level in these two countries:

(1) The policies (e.g., the status of EFL programmes at schools; the official starting age; curriculum objectives)

(2) The provisions (e.g., instructional time; teacher qualifications; resources available)

(3) The assessment/evaluation (e.g., formative and summative assessments)

Table 2.1 summarises the policies and related provisions in EFL learning at the primary level in South Korea and Japan. Key issues are discussed with respect to the information provided below.

Table 2. 1 Overview of the EFL curriculum policies and its provisions at the primary school level in South Korea and Japan

Countries	**Curriculum Policies**				**Provision**			**Assessment**
	Status of EFL program at the primary level	*Official starting age*	*Curriculum objectives*	*Expected level of achievement at the end of primary*	*Instructional time*	*Teacher qualifications*	*Materials available*	
South Korea	Starting from 1997, English is officially required in the curriculum as a compulsory subject as outlined by the Korean Ministry of Education.	Grade 3, Age 9-10	To develop basic communication skills; to develop students' interest and positive attitudes toward EFL learning.	There is a locally set examination or test with its own standards which are different from the Common European Framework.	Grades 3 and 4: two hours per week, totaling 29 hours per year. Grades 5 and 6: three hours per week, totaling 32 hours per year.	Generalist (i.e., homeroom) teacher; specialist teacher; native English speaker teacher	A series of textbooks approved by the government.	Assessment is required; the test or exam is provided by the national or local authority outside the school; classroom-based assessments are recommended.
Japan	Starting from 2011, English is officially required in the curriculum in the form of 'Foreign Language Activities' rather than a subject as outlined by the Japanese Ministry of Education, Culture, Sports.[1]	Grade 5, Age 10-11	To establish students' communication abilities in English; to deepen the experiential understanding of the languages and cultures of Japan and foreign countries	There is no stated level for children to reach by the end of primary school.	Grades 5 and 6, one hour per week, totaling 35 hours per year.	Generalist (i.e., homeroom) teacher; native English speakers with no formal teaching qualifications	Officially approved textbooks (known as 'English notes') used for 'Foreign Language activities'.	No requirement for formal assessment.

Note. This table is based on the following sources: Japanese Ministry of Education, Culture, Sports (2011); Korean Ministry of Education (1997); Rixon (2013); Butler (2015)

[1] According to Nemoto (2018), a new policy regarding the English education at the primary levels in state-funded schools in Japan will be officially implemented from April 2020. A number of changes will be made: (1) In Grades 3 and 4, the English subject will be taught through in a form of "Foreign Language Activities" with one hour per week, totalling 35 hours per year; (2) In Grades 5 and 6, English will be treated as a "formally assessed subject" with two hours per week, totalling 70 hours per year. A textbook which is officially approved by the government will be used. Students at this stage will be required to develop communicative skills as well as the basic literacy skills (e.g. reading and writing simples words and full sentences); (3) There will be an increased number of trained specialist English instructors available.

In terms of EFL policies currently implemented in East Asia, it is important to note that these two countries are making English compulsory at the primary level. Under the uniform national curriculum in South Korea, since 1997 English has been taught as a mandatory subject starting in Grade 3, with the central goal of developing oral communication skills and favourable attitudes towards English and English learning (Korean Ministry of Education, 1997). In Japan, the government has recently introduced English into the national curriculum in the form of 'Foreign Language Activities' for Grades 5 and 6, starting in the 2011 school year (Japanese Ministry of Education, Culture, Sports, 2011). The 'Foreign Language Activities' are not formally taught academic subjects (e.g., mathematics and history) but are regarded as communication activities with the aim of deepening students' understanding of the languages and cultures of Japan and foreign countries (Japanese Ministry of Education, Culture, Sports, 2011). In both countries, young EFL learners are not expected to achieve a native level of proficiency; instead, curriculum objectives emphasise the importance of communicative competence and of a favourable impact on L1 development.

With regard to the actual provisions for early EFL instruction in formal instructional contexts, there are some relevant factors that need to be considered: the time allocated to EFL learning; the qualifications of teachers and the resources available. In South Korea, generalist teachers[2] and specialist English teachers teach English as a regular subject for two hours per week in Grades 3 and 4 (i.e., for a total of 29 hours per year) and for three hours per week in Grades 5 and 6 (i.e., for a total of 32 hours per year). The textbooks used were developed and approved by the governments; they offer varied options to meet the diverse needs of students. In Japan, primary school students have English activities one hour per week (i.e., for a total of 35 hours per year) with their homeroom teachers and/or with native English speakers. These instructors do not have specific qualifications for EFL instruction (Butler, 2015). The materials were government-developed textbooks. Though the early EFL formal instruction in both countries starts at different points, there are some common characteristics regarding the actual provisions: (1) a fairly limited amount of time is dedicated to instruction per week; (2) teachers are generalist teachers rather than the English specialists, and teachers' proficiency in English often falls far short of a native speaker's level (Johnstone, 2009); (3) the materials are standardized and officially approved by the central government at the national level (Butler, 2015).

In terms of assessing the learning outcomes of young EFL learners, the Korean government requires a summative assessment with classroom-based assessments (e.g., formative assessments) strongly recommended. In Japan, teachers are not required by the government to assess their students by any formal measures. Although some teachers do use self-assessments (e.g., asking students to reflect after lessons on their performance), this assessment style does not provide sufficient information about

[2] The generalist teachers are also called homeroom teachers, often teaching according to the curriculum to their students.

students' learning progression and outcomes (Butler, 2008; Nikolov & Mihaljević Djigunović, 2011). The governments of both countries have not provided any formal systematic measurements for assessing young EFL learners' language proficiency in ways that directly match the objectives of their EFL programmes at the primary level.

2.3.2.2 The new National English Curriculum Standards and primary EFL provisions in China

In China, many of the recent developments in EFL learning at the primary level are parallel to those in South Korea and Japan. This section summarises China's recent policy change and its provisions for EFL instruction at the primary school level. The perceived difficulties in implementing the new national curriculum are also discussed.

Chinese primary school English curriculum reform

Prior to 2001, English was not compulsory in primary schools in China. The teaching of English at this level had not been standardised for multiple reasons, including the unrecognised position of English in the school curriculum, the limited availability of qualified teachers and the lack of appropriate teaching materials. However, in 2001, the Chinese Ministry of Education (MOE) issued a document called '*English Curriculum Standards for Compulsory Education and Senior High Schools (Trial Version)'*. With this document, English became a compulsory subject in primary education beginning in Grade 3 (ages 8-9). The new curriculum was to start in cities and then progressively move to counties, towns, and villages(Chinese Ministry of Education, 2001).

After ten years of implementing the trial version *Standards*, the Chinese MOE released the formal version, '*English Curriculum Standards for Compulsory Education*', in 2011 with minor revisions[3] (Chinese Ministry of Education, 2011). The new National English Curriculum Standards (NECS) emphasise that the overall objectives of the EFL program are "to develop students' comprehensive language ability of English, to facilitate children's mental and cognitive development and to develop affective attitudes and cultural awareness towards EFL learning" (Chinese Ministry of Education, 2011, p. 2). To that end, targets are categorized in five areas: language skills, language knowledge, affective attitudes, learning strategies and cultural awareness (Table 2.2). In terms of language skills, the new NECS includes very detailed targets in the can-do statement format regarding what students should achieve for each skill domain. These targets are divided into two levels for primary school students. Students are expected to meet Level 1 by Grade 4 and to meet Level 2 upon completion of Grade 6. For the remaining four categories, targets are set only for Level 2 and Level 5, which correspond to the end

[3] The minor revisions include removing the standards for senior high schools; setting targets in a coherent way from Grade 1 to Grade 9, through six-year primary to three-year junior secondary education.

of primary education (i.e., Grade 6) and to the end of junior secondary education (i.e., Grade 9).

Table 2. 2 The new NECS goals (Chinese Ministry of Education, 2011)

Target categories	Target subcategories	Target standards
Language skills	listening, speaking, reading and writing	Levels 1, 2, 3, 4, 5
Language knowledge	pronunciation, vocabulary, grammar, functional expression and daily topics	Levels 2 and 5
Affective attitudes	interest, motivation, confidence, determination, teamwork, nationalistic awareness and international perspectives	Levels 2 and 5
Learning strategies	cognitive strategies, management strategies, communication strategies and resourcing strategies.	Levels 2 and 5
Cultural awareness	cultural knowledge, cultural understanding, intercultural communication awareness and abilities.	Levels 2 and 5

The new national curriculum emphasises the importance of language skills independent of language structure and language knowledge and aims to balance literacy (i.e., reading and writing) and oracy (i.e., listening and speaking). The level descriptors provide teachers with more explicit guidelines on developing students' comprehensive competence in four main language skills at different learning stages. In addition, the communicative task-based language teaching approach is advocated by the NECS. This approach makes it clear that teachers should avoid purely focusing on the delivery of language structure and instead focus on designing authentic tasks from real life that require students to use English to communicate.

Primary EFL provisions in China and their perceived problems

The actual provisions of EFL education in China include instruction time, textbooks, teacher qualifications and the evaluation system, which are discussed below.

The instructional time for English, as required by the new NECS, is two times a week, totaling a minimum of at least 80 to 90 minutes per week. According to a recent survey (Rixon, 2013), the actual average instructional time allocated to English met only the minimum requirements of two class hours per week[4]; there was also a significant variation in instructional time allocation between urban and rural schools.

A series of official textbooks, which have been approved by the National Textbook Review Committee under the Chinese MOE, is formally recommended to schools throughout the country. Although these materials might accurately reflect the language and activities in major English-speaking countries, these

[4] One class hour is usually around 40 to 45 minutes. The exact duration varies in actual practice.

textbooks do not necessarily relate to "Asian students' daily lives or correspond to the kinds of language that they would use in real communicative contexts as a means of global communication" (Butler, 2011, p. 42).

For instructors, a certified qualification in English teaching is required and pre-service and in-service training courses for implementing the curriculum are provided. However, there is a shortage of qualified English teachers at the primary level. According to a recent survey of English teachers from seven provinces across mainland China, over 90% of respondents held degrees at or above the college level. However, among them, only 53.1% of respondents held degrees in English while nearly 40% had degrees in subjects other than English; the remaining 6.6% held secondary certificates (Wang, 2009). The situation of the percentage of respondents holding degrees in English or other subjects at or above the college level was even lower in developing districts (Luo, 2014).

As for methods of assessment, the new NECS recommends both summative assessment and formative assessment with learner-centred principles and a focus on assessing students' language use abilities (Chinese Ministry of Education, 2011). However, the reality of assessment and testing at the primary level in China is likely to vary from government expectations due to an unchanged examination-oriented educational system. For instance, a recent study that interviewed 73 teachers in China from state-funded schools, ranging from the primary to the senior secondary level, found that while all the teachers appreciated the recommended methods in the NECS on assessing students' overall development in English learning, they thought those methods were not in accordance with a national exam culture, in which both students and teachers are eventually evaluated based on the results of summative examination (Li & Baldauf, 2011). Thus, although Chinese primary students can presently enter junior secondary school automatically without any examination pressures, given China's system of free compulsory education for the first nine grades, students still face high-stakes examinations, such as Senior High School Entrance Examination and the National College Entrance Examination. These exams play a decisive role in which senior secondary schools and colleges students are permitted to attend (Li & Baldauf, 2011; Luo, 2014).

2.3.3 The challenging factors of curriculum implementation

As discussed earlier in section 2.3.1, the age factor may be an important variable for the learning outcomes of early EFL programs. However, after having reviewed current early EFL programs, policies and available provisions, this study has noted that there are other, arguably more important, contextual factors that interact with the age of acquisition. These other contextual factors include the quality and amount of input, the available instruction time and the support given to EFL teachers (Dekeyser & Larson-Hall, 2005; Murphy, 2014; Nikolov & Mihaljević Djigunović, 2006; Nunan, 2003). Therefore, in this section, these factors, which have been identified and frequently mentioned in prior studies in

terms of curriculum implementation, are discussed, with the goal of gaining a more comprehensive understanding of early EFL learning and its learning outcomes.

2.3.3.1 Teachers' language proficiency and professional knowledge

Teachers' language proficiency and their professional knowledge are discussed under one heading since both factors are mentioned simultaneously in prior studies. Both factors also characterize desired qualifications for effective teachers; that is, a qualified EFL teacher of young language learners should not only be highly proficient in both the first language and the target language but also should possess professional knowledge of the characteristics of young learners and of age-appropriate teaching methodologies (Butler, 2015). However, as has been revealed in prior studies, there is still a lack of qualified teachers, especially at the primary level. According to the results of a survey study conducted by Nunan (2003) across seven countries and districts in the Asia-Pacific regions, English teachers in all the surveyed countries and districts had low language proficiency. Their levels were insufficient to implement effective language teaching by providing learners with quality input and interaction. Given this finding, a self-evaluated survey of primary EFL teachers from Korea, Japan and Taiwan was conducted to investigate teachers' current perceived proficiency levels as well as the desired proficiency levels required for successful English teaching in their teaching contexts (Butler, 2004). The results indicated that, consistently, among the majority of participating teachers, there were substantial gaps between the teachers' current proficiency level and the minimal level required for teaching. Specifically, the widest gaps were found in the production domains, particularly in oral grammar. Teachers' low oral proficiency might lead to insufficient oral communicative skills in the classroom and so contribute to the teachers' lack of confidence in their ability to instruct students in speaking English (Butler, 2015; Hu & McKay, 2012). This low oral proficiency might also influence classroom-based assessments, since students might be unable to understand or to respond to a teacher's English oral instructions, questions and feedback. Therefore, this aspect was considered for further investigation in the current study.

2.3.3.2 Teaching resources

As was discussed in the previous section, teachers' lack of proficiency and professional knowledge demonstrated an urgent need for support from various sources (e.g., teacher trainings and teaching materials), at both the national and local levels. However, in many countries, such resources were not available because these countries took a short-term approach when adopting policies requiring the study of English as a compulsory subject. As a result, the curriculum standards that were developed were based on theoretical assumptions and lacked a basis in thorough investigation and empirical studies. These standards were also put into implementation in a top-down manner, especially in East Asian countries (Enever, 2018). Moreover, the associated pre- and in-service trainings were not always in

place; therefore, teachers could not fully understand the new policies and have necessary resources to implement it (Enever et al., 2009; Garton, Copland, & Burns, 2011; Hayes, 2014; Nunan, 2003). For example, one study revealed that, in China, although the government-approved textbooks were asserted to be compatible with national curriculum principles, and the local government at the provincial level was required to support the curriculum implementation, the actual trainings and seminars were delivered in a one-time, mass-lecture style; also, there were no further supplementary trainings or materials offered (Deng & Carless, 2009). Moreover, according to local teachers, the content of the in-service training consisted of theories that were not applicable to actual classroom practices (Hayes, 2014). As a result, there was little change in practice, and teachers reverted back to their original teaching concepts and methodologies (i.e., audio-lingual or grammar-translation methods) (Butler, 2015).

2.3.3.3 Classroom-level constraints

Classroom-level constraints were those related to contextual constraints at the classroom level, including limited instructional hours, large class size and the mixed abilities of students. Such issues have been frequently mentioned as main barriers for teaching and assessing students' oral and aural skills in class, especially in East Asian contexts (Butler, 2011; Garton et al., 2011; Su, 2006). Specifically, the limited amount of time dedicated to instruction per week resulted in limited input and in limited opportunities for students to take part in oral-production and task-based activities. Limited instructional time also challenged teachers to cover the content of the textbook within the available timeframe. Similarly, a large class size made it difficult and time-consuming for teachers to achieve universal student participation in oral English tasks (Butler, 2005). In addition, large class size also posed challenges for teachers' ability to monitor students' language use while also maintaining 'classroom harmonization' (Butler, 2005). Therefore, given such classroom-level constraints, the present study also considers which classroom-based assessment activities were used and what criteria teachers used to assess their students' learning outcomes

2.3.3.4 Exam washback effects

As was summarized in Section 3.2.2.1, although a primary goal of government policies regarding compulsory EFL education in East Asian countries was to develop students' communicative competence, the examination-oriented educational systems did not change. As a result, even though teachers were being encouraged to use communicative task-based teaching approaches to enhance students' oral skills, it was likely difficult for these instructors, tasked with preparing students for formal public exams (which were oriented toward the assessment of grammar, not oracy, and were completed in paper-and-pencil format), to introduce more interactive oral activities (Enever et al., 2009; Garton et al., 2011). The tendency for teachers not to implement such activities can be labelled, with other

recidivist habits, as "exam washback effects". These perceived effects were pervasive, but likely varied greatly among teachers. For example, a recent case study of primary EFL teachers in China used classroom observations and in-depth interviews to investigate the impact of exams on classroom pedagogy (Deng & Carless, 2010). The results revealed that an individual teacher's teaching philosophy and opinions about assessment played an important role in the perceived influence of examinations on that teacher's classroom pedagogy. School differences, which may be defined as school's orientations towards the examination preparations or innovative teaching methods (e.g., communicative teaching method), also were an important factor.

2.4 Assessment of young EFL learners

Apart from the curriculum policies, as well as the actual provisions and perceived difficulties discussed above, there is still little known today about the learning outcomes of early EFL programs at the primary school level in East Asian countries. Butler (2009) summarized the reasons for such a lack of literature regarding East Asia. First, the introduction of early EFL programs into the formal school setting at the primary level began just at the turn of the 21st century and collecting substantial information on the effects of EFL programs requires long-term study. Second, there is still a lack of reliable and effective assessments to evaluate students' attainment in line with the goals stated in the curriculum, both for internal use by teachers and for external use by local governments and parents (Butler, 2009a). Due to the importance of assessing student learning outcomes, tracking learning progression and diagnosing weaknesses and strengths, the assessment of young EFL learners has gradually gained attention from both governments and researchers in East Asia (Butler, 2015; McKay, 2006). In addition, although oral language develops in the early stage of EFL learning and is seen as the core competence in young learners' curriculum, oral language is seldom assessed in external public exams; as a result, we have little knowledge about students' actual learning outcomes in the oral skills domain (McKay, 2006). Therefore, in the following sections, assessments of listening and speaking are discussed first. This discussion is then followed by a review of assessments for internal and external use.

2.4.1 Constructs of language ability in relation to assessments in listening and speaking

Developing EFL communicative competence is a frequently emphasised goal in the national curriculum policies of East Asian countries. China is no exception. As required by the recent NECS, one of the main objectives for China's primary EFL program is to encourage students to use English to communicate and to solve problems. To investigate the current Chinese EFL primary students' oral and aural language skills in relation to the targets that are stipulated in the NECS, it is essential to define the constructs of language ability or proficiency in a useful way.

Language proficiency was initially described as a "skills-and-components" model, composed of four

basic skills, including listening, speaking, reading and writing, and three components, including pronunciation, grammar and vocabulary (Carroll, 1968; Lado, 1961). However, this model did not consider language use or real-world contexts, and thus was eventually rejected by practitioners. A later framework, which was proposed by Bachman and Palmer (1996, 2010), defined language proficiency from a more pragmatic perspective as the "ability to use the language for the purpose of achieving a particular goal or objective in a particular situation" (Bachman & Palmer, 1996, cited in McKay, 2006, p. 27). This framework was chosen by McKay (2006) to describe children's ability to use language communicatively because it provided a relatively comprehensive tool to understand the nature of language ability at two levels: (1) it addresses children's creative using language for the purpose of interaction (e.g., to respond to teacher's questions); and (2) it also takes children's immediate physical environments (e.g., when they are discussing a topic with their peers in a group) and needs into consideration, and it considers these factors as equally important to meaningful communications.

This framework also defines concepts such as *language knowledge*, *strategic competence* (i.e., metacognitive function) and other *individual attributes* (e.g., personal attributes, topical knowledge and affective schemata) in ways that address to language use in real-world contexts. For example, *language knowledge* is further divided into two subfields: one structural, the other pragmatic. The first subfield addresses how utterances are organised (i.e., organisational knowledge), with a focus on vocabulary, syntax, phonology and cohesion; but the second addresses how utterances are related to the communicative goals of the language user (i.e., to pragmatic knowledge), which in turn touches on two areas: functional and sociolinguistics knowledge (Bachman & Palmer, 1996, 2010). This attention to the detailed and organised components of language use in real-world contexts aligns with the emphasis of the NECS on developing students' communicative language use competence. It also addresses expected achievement within the scope of functions, much as NECS targets do (e.g., one listening NECS target at Level 2 is "can understand common instructions and directions and respond properly").

This framework's attention to the individual attributes of young learners (namely, to their personal attributes; their topical knowledge; and their affective schemata) suggests that assessments of EFL learners should also take each of these three types of individual attributes into consideration. For example, as was discussed in Section 2.2.2.1 children undergo a period of cognitive, social, physical and emotional maturation, and their (developing) personal attributes influence both their approach to and performance of language assessment tasks. Therefore, in the current study, it will be important to investigate teachers' assessment practices as well as EFL students' assessment performance at different grade levels (i.e., lower and upper grade levels). Specifically, whether the teachers' assessment practices are responsive to children's personal attributes or changing development at different grade levels and whether students at different grade levels show different affective attitudes towards the same assessment activity.

It will also be important to consider children's topical knowledge. As children use a target language to communicate, they draw on their background and topical knowledge. At the same time, as the children grow up, they are also developing their knowledge of the world, which is said to advance through four conceptual spheres: direct experience, which can be seen, heard and touched; daily or personal experience; indirect experience, such as literature or imagination; and, lastly, abstract topics, such as social or environment issues (Scarino, Vale, McKay, & Clarks, 1988). Therefore, the selection of appropriate assessment task should consider students' current topical knowledge and their current conceptual spheres.

Third and finally, the framework suggests that is also important that decisions made regarding assessments consider the influence of children's, affective schemata upon results. As suggested by Bachman and Palmer (1996), children are also maturing emotionally. The topics, materials of assessment, and even their familiarity with the assessors could influence their emotional responses, affecting their performance on assessments. Moreover, because children, as they grow, gain greater confidence, independence, and collaborative skills, they will also likely exhibit different affective attitudes at different ages toward the same assessment task. Therefore, in order to bring primary EFL students' best efforts into their performance during assessment, it is important to provide a safe and supportive assessment environment that can promote their feelings of comfort based on their age or grade level. It is also worth investigating primary EFL students' affective attitudes towards classroom-based assessment activities at different grade levels.

Therefore, in the current study, relying on Bachman and Palmer's (2010) framework, taking into account both language skills and students' affective attitudes at two key grade levels, not only are students' listening and speaking skills are assessed by a series of assessment instruments, but these skills are also considered alongside the students' motivation, confidence and anxiety towards language learning.

2.4.2 Assessing the process: classroom-based assessment

This section begins with an overview of prior definitions of classroom-based assessment. First, the multiple dimensions and functions associated with such assessment are discussed; then, classroom-based assessment is defined for the current study. Empirical studies related to classroom-based assessment in the early EFL context are also reviewed and discussed.

2.4.2.1 Defining classroom-based assessment

Although classroom-based assessment has been widely used by different researchers in the existing literature, the term has been defined variously, from different dimensions, with many alternative uses,

focused on different aspects of assessment. As a result, the literature presents the following variety of assessment terms: *alternative assessment*, *teacher-based assessment*, *formative assessment* or *assessment for learning*. In this section, the development of these definitions in recent literature is summarised and discussed.

In the late 1990s, drawing on a review of about 250 studies focused on assessment conducted in the classroom, Black and Wiliam (1998a)'s seminal paper exhorted researchers to investigate what happened within the classroom. The paper regarded the classroom as essentially still a *black box*, since researchers had little data regarding teachers' daily assessment practices, even though these practices, the paper concluded, held the potential to increase students' achievement. According to Black and Wiliam (1998a)'s definition, classroom-based assessment was to be regarded as formative assessment, that is, as "all activities undertaken by teachers – and by their students in assessing themselves – that provide information to be used as feedback to modify teaching and learning activities"(Black & Wiliam, 1998b, p. 8). These findings indicated that the role of assessment was not limited to measurements of students' learning by psychometric instruments; rather, assessment itself could be viewed as a formative approach to improve teaching and enhance learning.

Rea-Dickins and Gardner (2000) further broadened the functions of classroom-based assessment by embracing both summative and formative purposes in their study. This study, which used self-reported interviews and classroom observations of 28 teachers in a primary school language learning context, found that the same classroom assessment task could serve either summative or formative purposes; in fact, the distinction between the two purposes was not easily identified. Rea-Dickins (2001) later defined classroom-based assessment in a comprehensive manner, emphasising that the term "covered the full range of elicitation techniques that were used by teachers in their classrooms" (Rea-Dickins, 2000, p. 116) and that it was "routinely embedded within good classroom practices" (Rea-Dickins, 2001, p. 434). This finding meant that classroom-based assessment could be treated as an integral part of effective teaching. In addition, Rea-Dickins (2001) further argued that the teacher was not the only agent of assessment; during the assessment process, learners could also function as assessors.

In more recent literature, Davison and Leung (2009) defined classroom assessment in terms of teacher-based assessment, prefacing the term with a number of contextual features(e.g., "teacher-mediated, context-based and classroom-embedded assessment practice") that contrasted with the "externally set large scale formal examinations used primarily for selection and/or accountability purposes" (Davison & Leung, 2009, p. 395). Their paper also proposed a framework for effective teacher-based assessment in the classroom. According to this framework, there are four key steps that are incorporated into the assessment cycle: plan assessments; collect information about student learning; make professional judgements and provide appropriate feedback or advice. Davison and Leung (2009) further pointed out

that these key steps were not arranged at the end of teaching; rather, they were "continuous and naturally embedded into every stage of the teaching and learning cycle" (p. 399). In addition, similar to the findings of Rea-Dickins (2001), Davison and Leung (2009) also indicated that there was no clear distinction between formative and summative assessment; instead, they placed teacher-based assessment on a continuum with four possible variations, ranging from formal planned summative assessment to informal contingent assessment-while-teaching. All of these assessments were treated as formative when the results were used for improving student learning.

It is worth noting that Wiliam and Thompson (2007) had previously proposed a framework (Figure 2.1) for effective formative assessment. Their framework incorporated five classroom strategies (numbered in Figure 2.1), which were said to be involved in three key processes of teaching and learning (the first row in Figure 2.1) across different agents (i.e., teacher, peers and learners). According to their framework, learning intentions are effectively realized as the final outcomes through these five strategies. Black and Wiliam (2009) further elaborated upon this framework and drew researchers' attention towards "moments of contingency" within formative assessment, which they treated as the central feature distinguishing these assessments from ongoing teaching and learning. Such moments of contingency, as suggested, were the real-time formative interactions that occurred in the form of teachers' feedback regarding the productions made by individual learners, a group or the whole class. These formative interactions followed the format of Initiation-Reply-Evaluation, or IRE (Mehan, 1979), which represented the basic steps involved in the implementation process.

Figure 2. 1 Wiliam and Thompson's framework of formative assessment in class

<table>
<tr><th>Agent</th><th>Where the learner is going</th><th>Where the learner is right now</th><th>How to get there</th></tr>
<tr><td>Teacher</td><td rowspan="3">Clarifying learning intentions and sharing and criteria for success (1)</td><td>Engineering effective classroom discussions, activities and tasks that elicit evidence of learning (2)</td><td>Providing feedback that moves learners forward (3)</td></tr>
<tr><td>Peer</td><td colspan="2" rowspan="2">Activating students as instructional resources for one another (4)</td></tr>
<tr><td>Learner</td><td colspan="2">Activating students as the owners of their own learning (5)</td></tr>
<tr><td></td><td>Understanding learning intentions and criteria for success (1)</td><td colspan="2"></td></tr>
</table>

Based on an empirical study of students in Year 6 and Year 7 who were learning Indonesian as a foreign language in two Australian schools, Hill and McNamara (2012) proposed a comprehensive framework for the classroom-based assessment process. This framework provided detailed and structured dimensions for researchers to investigate classroom-based assessment based on the following factors: how teachers plan and frame the assessment; how teachers actually conduct the assessment and how

teachers finally make use of the assessment data. In addition, this framework also took into account both teachers' and learners' beliefs about learning and assessment. This approach is also adopted in the current study since it is important to understand teachers' actual enactment of classroom-based assessment together with the teachers' and their students' opinions.

To summarize, according to prior definitions and theorisations of classroom-based assessment, a number of common features can be captured and implied for the current study. These features are teacher-mediated, classroom and context based and can serve for both summative and formative purposes, with the aim of facilitating students' achievement of curriculum goals. Though integrated with daily normal teachings, classroom-based assessment can be captured through teachers' and students' formative interactions in the IRE format.

2.4.2.2 Empirical studies on classroom-based assessment in the early EFL programmes

Of the empirical studies that have been conducted on young language learners in the early formal instructional setting, those focusing on classroom-based assessment are limited.

Some of the studies focused on investigating the procedures of effective classroom-based assessment. These studies were reviewed and discussed in Section 2.4.2.1 (Hill & McNamara, 2011; Rea-Dickins & Gardner, 2000).

Other studies focused on examining the effectiveness of specific aspects of classroom-based assessment. For example, Butler and Lee (2006) investigated the validity of employing self-assessment in the primary classroom in South Korea. The study found that students could evaluate their own performance more accurately through on-task self-assessment (i.e., assessment in which students self-evaluate themselves immediately after accomplishing a task or test) than off-task self-assessment. In addition, the study found that older learners (i.e., Grade 6) could assess themselves more accurately than younger learners (i.e., Grade 4) could. The authors' later study also confirmed that young EFL learners could develop accuracy in self-assessment over time and that the administration of regular self-assessment could positively influence students' language learning (Butler & Lee, 2010). Another study by one of the same co-authors, focused on the primary level in China, investigated whether students' performance could be better evaluated through task-based paired-assessment (i.e., peer interaction) or task-based individual-assessment (i.e., student-teacher interaction). The results revealed that although younger learners benefited from the individual-assessment format, paired-assessment could yield effective language use with communicative strategies if students were carefully guided and trained to collaborate with each other (Butler & Zeng, 2014). These studies, however, were more or less treated as intervention studies. They did not reflect teachers' actual practices of classroom-based assessment. The effectiveness of introducing these assessment techniques (e.g., on-task self-assessment) also did not take contextual

factors into consideration. Doing so is important, especially in the East Asian classroom context, which usually involves a large class size, a limited number of instructional hours and a shortage of qualified EFL teachers.

In terms of actual classroom-based assessment practices, only a limited number of studies have been conducted at the primary level, and most of these were large-scale survey studies. For example, an international survey study was conducted by Rea-Dickins and Rixon (1999) to investigate 120 EFL teachers' practices and perceptions of the assessment of young language learners. The results of this survey study revealed that the most frequently used type of classroom-based assessment was the traditional paper-and-pencil test focused on the analysis of grammatical and lexical items. The researchers further pointed out that such tests did not assess students' language use ability and that components related to students' listening and speaking skills were also often neglected. Using the adapted survey of the previous study, Brumen, Cagran and Rixon (2009) investigated teachers' practices and beliefs about the assessment of young EFL leaners in three Eastern European countries. A number of characteristics regarding teachers' assessment practices were captured. The results revealed that EFL teachers were aware of the importance of speaking and listening skills to language learning at the primary level and that the most frequently used assessment type was the oral interview. However, the top 2 and top 3 most frequently used assessments were still traditional paper-and-pencil tests, which either were developed by the teacher or were taken directly from the textbook. In addition, this study investigated the actual assessment activities that teachers used during class. The results indicated that the most frequently used ones were less cognitively demanding and were presented in a controlled and guided way (e.g., gap-fill activities and oral interviews). A similar study conducted by Chan (2008) in Taiwan addressed some 520 primary EFL teachers' beliefs, perceptions and practices regarding the assessment of young learners. The results of the study revealed that, although teachers acknowledged the importance of employing multiple types of classroom-based assessments, due to time constraints and large class sizes, as well as the lack of adequate professional guidance to implement assessments for formative purposes, the paper-and-pencil tests and classroom observations were mostly unutilized by the teachers. However, due to the large-scale nature of this quantitative study, it did not detail the content of the paper-and-pencil tests. The study also did not elaborate upon the methodologies by which teachers formatively observed students during or the means by which teachers utilized their observations to inform their teaching or to facilitate student learning. Taken together, although the previous large-scale survey studies provided some insights regarding teachers' perceptions towards classroom-based assessment in the primary context, and although these studies found a similar tendency among teachers to rely on traditional paper-and-pencil tests, little is really known about either the detailed associated features or the actual implementation of commonly used classroom-based assessment activities, especially in relation to set curriculum goals or targets.

2.4.3 Assessing the outcomes

The previous section conveyed the importance of studies investigating the types of assessments used in the classroom to inform daily teaching and learning. However, there is also a need to quantify what students can do in terms of their proficiency outcomes. Nonetheless, at the present time, it is still a challenge to find appropriate instruments to assess EFL children's foreign language outcomes or to map these learning outcomes with the targets set by national curriculum standards. As was discussed in Section 2.2.2.1, young language learners experience a period of cognitive, social, emotional and physical growth, and unlike adults, they have a particular vulnerability that requires careful attention. These young language learners are also developing their literacy skills in both their first language and English (McKay, 2006). These special characteristics of young EFL learners need to be considered when selecting appropriate assessment tasks and procedures. In this part, the first section discusses and evaluates three types of externally used assessments: standardized large-scale tests, non-standardized tests (which are administered for research purposes) and tests developed by local educational authorities. The second section mainly reviews current studies focused on assessing young learners' non-linguistic learning outcomes, that is, their affective attitudes towards EFL learning.

2.4.3.1 Assessing linguistic outcomes: the available assessment instruments

Large scale standardized tests

Large-scale tests are usually defined as "standardised externally designed tests which follow a consistent set of procedures for designing, administrating and scoring" (McKay, 2006, p. 315). Such tests are regarded as being credible since they are developed through a series of uniform and standardised test development procedures (e.g., design phase, operationalisation phase and administration phase) and with careful attention to the tests' validity and reliability (McKay, 2006). The results of these tests are also considered to be reliable and valid and can be compared between schools locally and globally. Moreover, large-scale tests are usually administered in the pencil-and-paper format. A benefit of employing such tests is that they are less costly, easy to administer and time-saving.

However, administering large-scale tests to young language learners has also elicited criticism for several reasons. First, large-scale tests cannot provide timely feedback for promoting effective teaching and learning. Instead, such tests belong to the category of administrative tests, which are used to inform parents, schools and local education authorities about children's current learning stage in relation to other children's. Second, large-scale standardised tests do not assess children's language knowledge in a natural setting; rather, they assess children in an exam context, in which children are asked to answer questions and to respond to items on paper. Despite these objections, some researchers argue that large-scale tests can have motivational value in the language education of young learners if the tests are

developed appropriately, taking the characteristics of young learners into account (McKay, 2006).

The Cambridge Young Learners English (YLE) tests, developed and administered by the University of Cambridge ESOL Examinations, is one such battery of large-scale tests. This set of tests was designed to assess general English proficiency among young EFL learners ages 7 to 12 and is offered at three ability levels (Starters, Movers and Flyers). A review of the Cambridge YLE test indicated that the task items were not designed to be test-like but complied with the special needs and characteristics of young EFL learners. In other words, the tests were found to employ authentic and school-like materials and to use age-appropriate daily language; to aim to make the testing condition as anxiety-free as possible and to deliver the tests in an enjoyable and motivating way (Bailey, 2005). Therefore, the current study considers the Cambridge YLE tests as one of the instruments to assess EFL children's language ability during the pilot phase.

There also have been a series of individually administered standardised batteries developed for use in specific contexts and usually norm-referenced. For example, the British Ability Scale (BAS) has long been established as a leading standardised battery for use in the UK to assess children's overall educational achievement, cognitive ability and intellectual functioning across a wide age range from age 3 to age 17 years, 11 months (Elliot, Smith, & McCulloch, 1996). The BAS tests were designed for use by school teachers, clinical psychologists and researchers to assess children who might have learning or behavioural difficulties. The tests also provided a set of comprehensive normative scores, such that the results were reported or interpreted with reference to the achievement of other age-equivalent children. Although the BAS tests were norm-referenced in the UK and have not been standardised for the EFL population, they can still be valuable assessment instruments for EFL children, since these batteries include well-developed tasks that tap the essential components of specific language ability (e.g., verbal comprehension in BAS).

Non-standardised tests for research purpose

In addition to large-scale standardised tests, researchers have themselves designed specific tasks that are catered to the particular needs of young EFL learners (e.g., assessing listening or speaking).

Some researchers have used or designed specific tasks to elicit young EFL learners' language sample or to assess their language use (Szpotowicz, 2012; Szpotowicz & Lindgren, 2011). For example, in a large-scale, three-year longitudinal study of Early Language Learning in Europe (ELLiE), which aimed to evaluate early EFL programs in instructional contexts across seven European countries (Croatia, England, Italy, the Netherlands, Poland, Spain, and Sweden), researchers designed specific classroom-based reading, listening and focal learner speaking tasks to assess ELLiE learners' productive and receptive skills (Enever, 2011). As young learners have limited language proficiency, as well as limited

contact with foreign languages outside of school, the tests in ELLiE were designed as classroom tasks using authentic classroom language to ensure that the tasks were closely linked to the children's classroom experiences. The tasks were also designed to be both familiar and appealing to children. For example, in reading, selected-response tasks were used, where children were required to read a comic strip and complete the storyline by filling in empty speech bubbles from a selection of sentences. Such selected-response tasks, as noted by McKay (2006, p. 100), were appropriate to use for young learners since children's language participation in such tasks "involves a degree of spontaneity and creativity, they make their own meaning, producing meaning or comprehending meaning, according to the purpose and the requirements of the situation". However, the study also found that such tasks, which were designed in accordance with a specific syllabus or set of research needs, might not be equally appropriate for EFL children from different countries. Even though the same tasks were administered to children across seven participating countries, with the same tasks used to ensure the consistency of the test administration, the researchers of this study were aware of the difficulties in collecting reliable data with young EFL learners across participating countries. They recognised that, because of the focus early EFL programs, teaching approaches and learning contexts of early EFL programs varied significantly, children in some countries might find the same tasks challenging whereas children from other countries might find the tasks similar to those routinely used in their classrooms. Therefore, designing or finding an appropriate task to assess young EFL learners' language ability is still a challenging process.

Other tasks, though unstandardised, have also been widely used for research purposes across countries and with different population groups. For example, narratives have been elicited through an extensively studied storytelling task that uses a wordless picture book, such as *Frog, where are you?* (Mayer, 1969) or *Frog goes to dinner* (Mayer, 1974). This task has been used widely in a number of studies conducted with both children and adults. The popularity of this task likely lies in its ability to elicit long narratives involving a main storyline as well as a series of complicated plots, giving researchers an extensive narrative sample for further linguistic analysis (Isbell, Sobol, Lindauer, & Lowrance, 2004; Kang, 2012; Nicolopoulou, 2008). For these reasons, such a task is considered for use in this research for eliciting young EFL learners' narrative samples.

Tests by local education authority

Local governments have also developed external tests to assess students' English performance and to monitor students' learning progress year by year. In some East Asian countries (e.g., China, Korea and Malaysia) where testing pressure is high, young EFL learners are confronted with various mid-term and final examinations each year. These paper-and-pencil-based tests are generated from rubric-based English item banks, and are delivered in the form of discrete-point assessment (e.g., multiple-choice

items). The tests are also mostly concerned with measuring language knowledge (e.g., vocabulary and grammar), rather than language use ability (Butler, 2009b). Children's performance is evaluated on the basis of their scores. Such examinations have led to a negative washback effect since teachers are under pressure to prepare grammar-based teaching syllabi and to rehearse test-taking skills (Garton et al., 2011). Additionally, the results of these locally based external tests are not usually available to the public, and thus the detailed information or statistics collected from these tests still remains unknown (Butler, 2009b).

2.4.3.2 Assessing non-linguistics outcomes: students' affective attitudes

As was discussed in Section 2.2.2.2, young learners mature emotionally while they are learning language at the primary level. They develop their value system, and they incorporate their experience of the world, their motivations, interests and self-concepts into their language learning (McKay, 2006). This value system, which is also known as an appraisal system, is believed to strongly influence the children's language learning (Schumann, 1997). On the other hand, fostering young learners' positive affective attitudes is also considered as one of the aims, as well as a part of the explicit learning outcomes in some EFL programs (Mihaljević Djigunović, 2015). Therefore, in this section, prior studies investigating young learners' affective attitudes towards language learning are reviewed and discussed.

Only a few studies have examined changes, over time, in the affective attitudes of young language learners; this lack of studies is even greater for language learners at the primary level. This lack of research might be attributable to the commonly held, default belief that all young language learners hold positive attitudes towards language learning. However, such a belief has been gradually proved to be a misconception. The ELLiE project, a comparative analysis of data on young language learners' attitudes, motivation and perceptions of learning across seven European countries, indicated that, at the initial stage, the majority of young language learners displayed high motivation and positive attitudes towards foreign language learning. After three years, however, although most of the students retained positive attitudes, a slow decline in motivation was noted (Mihaljević Djigunović & Lopriore, 2011). Another study of ELLiE participants in Croatia further revealed that such a decline might be due to the students' less-than-ideal learning conditions, including the large class size, limited instructional time and unqualified teachers (Mihaljević Djigunović & Letica Krevelj, 2010). Related studies have identified similar results in East Asia. For example, a large-scale survey study conducted on Grade 3 and Grade 6 primary students in Japan revealed that there was a steady developmental decline in students' intrinsic and extrinsic motivations for EFL learning at school (Carreira, 2006). The potential reasons for this decline in young EFL learners in East Asia resembled those also suggested for learners in Europe: namely, a generalised decline in motivation in other subjects; inappropriate teaching methods and materials; and the absence of a motivating learning environment. In addition, young

learners with positive affective dispositions were found to show better performance on aural and oral production tasks (Mihaljević Djigunović & Lopriore, 2011).

A similar misbelief regarding anxiety has likely contributed to the dearth of research regarding the effects of anxiety on young language learners, especially at the primary level (Nikolov & Mihaljević Djigunović, 2011). This misbelief holds that young language learners are relatively free of anxiety during language learning at school. However, recent studies have revealed that anxiety can be observed in young language learners as well. The results indicated that as young language learners progressed to the upper grade levels, they experienced more anxiety (Mihaljević Djigunović, 2009). In addition, the students' level of anxiety was negatively correlated with the students' willingness to interact and communicate (MacIntyre, Baker, Clément, & Donovan, 2002), as well as with students' performance in listening comprehension tasks (Mihaljević Djigunović & Legac, 2008).

Since young learners' affective attitudes were shown to be related to these students' performance in the aural and oral related tasks, it might be implied that the selection of assessment tasks, as well as any associated implementation practices, should offer opportunities both to promote students' positive affective attitudes and to facilitate learning. It is therefore worthwhile in the current study to use the affective constructs (i.e., motivation, anxiety and confidence) mentioned above to assess primary students' general affective attitudes towards EFL learning in the instructional setting in China, for which there is still a lack of evidence. Moreover, the current study also investigates students' affective attitudes towards oral and aural assessment activities in class to provide further insight into students' attitudes towards specific activities, as well as to document at two key grade levels (i.e., Grade 3 and Grade 6) any observable changes in students' attitudes toward language learning. In addition, since most of the prior studies have been questionnaire-based, the current study has opted to use both a questionnaire format and student focus-group interviews to let students' voices be heard.

2.5 Summary of chapter two

The literature presented in this chapter first discussed the special characteristics of young EFL learners. It then reviewed current early EFL education policies and their implementation (with an emphasis on evaluating policy changes, provisions and their implementation) in several East Asian countries (e.g., Japan and Korea). The countries selected that are neighbouring countries to China that share with China similar socio-historical contexts in EFL learning and teaching. All of these countries have begun introducing English as either a compulsory subject or a foreign language activity into primary school curricula. However, as the literature review discussed above, large class sizes, limited instructional hours, the shortage of qualified English teachers and the influence of an exam-oriented assessment culture have been recognised as major obstacles for effectively implementing early EFL education curriculum at the primary school level in East Asian countries. In China, even as the new NECS, which

was recently released in 2011, has emphasised developing students' EFL communicative competence, due to the perceived difficulties and the lack of a standardised assessment system, both the classroom-based assessment methods utilized by teachers to assess Chinese young EFL learners in the formal instructional context and the effectiveness of the young EFL learners' learning outcomes (namely, whether they meet set targets) still remain unknown. The assessment of young language learners is, therefore, the primary focus of the current study. This assessment was discussed in the last section of this chapter, which reviewed both assessment of the learning process and assessment of learning outcomes. These two levels correspond to the discussion of classroom-based assessment and of externally used assessment instruments, respectively. In light of the literature reviewed, the current study has been informed by existing definitions of classroom-based assessment, but this study also has identified gaps in knowledge from prior empirical research. Thus, the chapter also tried to identify potential available instruments that could be used to assess young language learners.

Chapter 3 Research methodology

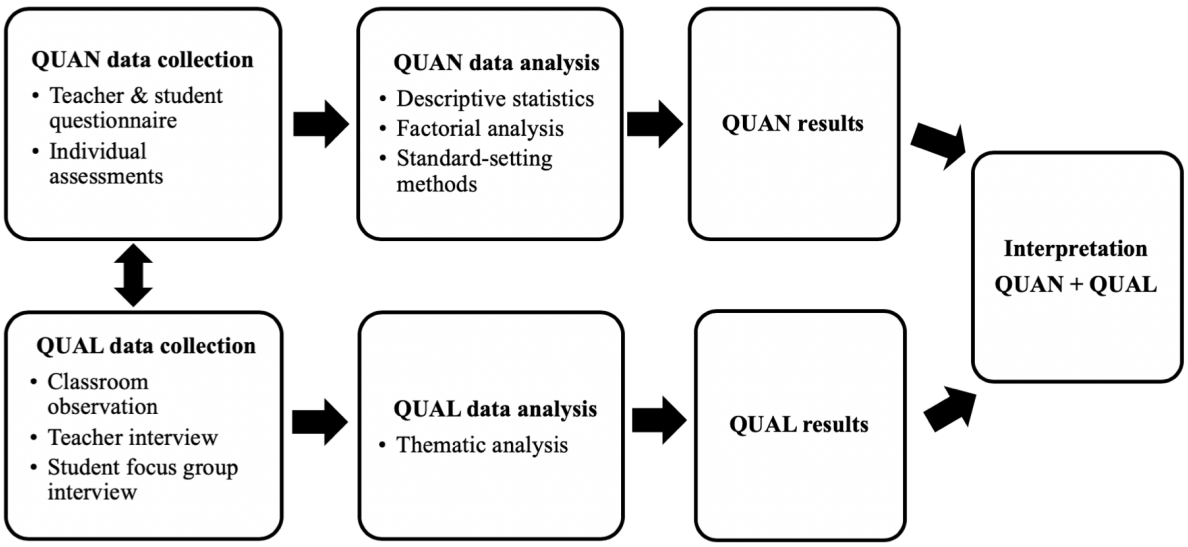
3.1 Introduction

This chapter first describes the research rationale and the mixed-methods approach, and gives an overview of the research design, explains the rationale for sample selection and gives profiles of participants. Then a detailed account of the rationale and adaption of each instrument used for the main study along with the pilot study results is provided. Additionally, a detailed explanation of data collection procedures data analysis for each instrument is presented. The chapter ends with a discussion of ethical considerations.

3.2 Research rationale and mix-methods research design

To get a rich picture of the actual EFL assessment practices and outcomes in relation to the national curriculum in Chinese classrooms, this descriptive and exploratory study used a mixed-methods research design with both qualitative and quantitative approaches (Mackey & Gass, 2015). The rationale for adopting such an approach is as follows. First, this choice was driven by the research questions, which were best served by a mixed-methods design that utilises the advantages of quantitative and qualitative research methods (Creswell & Clark, 2007). Second, methods and data were triangulated to cross-check; to allow for the comprehensive understandings and to increase the conclusiveness of the research findings (Creswell & Clark, 2007; Lindgren & Enever, 2017). For example, in the current study, the classroom observation enabled the researcher to better understand the context as well as to uncover the issues that might not be evident from the questionnaire and interview. Therefore, a convergent parallel model in the mixed-methods research design was used.

In the current study, quantitative data was obtained via 1) questionnaires from teachers and students and 2) students' performance in researcher-administered individual tests. The questionnaires were used to gather a range of information on attitudes towards curriculum implementation in teaching and assessment practices. The researcher-administered tests aimed to assess EFL primary students' actual listening and speaking skills. Qualitative data was collected through classroom observations and classroom stakeholder interviews, which provided in-depth information and details about what was happening in the classroom setting. Figure 3.1 illustrates the convergent parallel mixed-methods research design for the current study. It should be noted here, there was a two-way arrow inserted between the quantitative and qualitative data collection frames. It indicated that the construct of the questionnaire would direct the researcher's focus during the classroom observations and teacher and student interviews. At the meantime, the design and revision of the teacher and student questionnaire was also influenced by the classroom observations as well as teacher and student interviews.



Note: QUAN and QUAL stand for quantitative and qualitative research methodology respectively.

Figure 3. 1 Convergent parallel model mixed-methods design (Creswell & Clark, 2007)

A detailed account of the research activities, and the data each activity yielded, is summarized in Table 3.1.

Table 3. 1 Summary of research questions, participants and instruments

Main research questions	Sub research questions	Participants	Research instruments
Part I: Classroom context			
RQ1. How do EFL teachers from Grade 4 and Grade 6 conduct their assessments in listening and speaking in the classroom context in relation to the NECS targets?	RQ1.1 What classroom-based assessment activities do EFL teachers use for students in Grade 4 and Grade 6?	Primary EFL teachers in local district Four primary EFL teachers and their Grade 4 and Grade 6 students from School A and School B	Teacher questionnaire Classroom observation: The adapted COLT scheme Teacher interviews
	RQ1.2 How are these assessment activities implemented in the classroom?	Primary EFL teachers in local district Four primary EFL teachers and their Grade 4 and Grade 6 students from School A and School B	Teacher questionnaire Classroom observation: The adapted COLT scheme Teacher interviews
Part II: Teachers' perspectives			
RQ2. How is the intended NECS being interpreted by local EFL teachers?	RQ2.1 What are teachers' general perceptions of the curriculum implementation in aural and oral teaching?	Primary EFL teachers in local district Four primary EFL teachers and their Grade 4 and Grade 6 students from School A and School B	Teacher questionnaire Teacher interview
	RQ2.2 To what extent do teachers believe that their students will reach the curriculum targets in listening and speaking skills by the end of Grade 4 and Grade 6?	Primary EFL teachers in local district Four primary EFL teachers and their Grade 4 and Grade 6 students from School A and School B	Teacher questionnaire Teacher interview

Table 3.1 Continued

Main research questions	Sub research questions	Participants	Research instruments
Part III: Students' outcomes			
RQ3. According to the targets set by the NECS, what can young EFL learners actually do in speaking and listening upon finishing Grade 4 and Grade 6 respectively?	RQ3.1 What are Grade 4 and Grade 6 EFL students' performance on the two individual assessments, the BAS-II verbal comprehension and the Frog storytelling task?	Four classes of Grade 4 and Grade 6 students from School A and School B	Individual adapted assessments: BAS-II verbal comprehension Frog storytelling task
	RQ3.2 How do a group of teachers interpret the Grade 4 and Grade 6 students' results on the two individual assessments in relation to the levels (Level 1 and Level 2) in the NECS?	A panel of local teachers	Standard setting methods: The modified Angoff method The contrasting group method
RQ4. What are the attitudes of young EFL learners towards the learning, teaching and assessment of EFL in Chinese primary state schools?	RQ4.1 What are the affective attitudes of young EFL learners at two key grade levels (i.e. Grade 4 and Grade 6) towards the EFL learning?	Four classes of Grade 4 and Grade 6 students from School A and School B	Student questionnaire
	RQ4.2 What are young EFL learners' preferences for classroom-based assessment activities in listening and speaking at two key grade levels (Grade 4 and Grade 6)?	Four classes of Grade 4 and Grade 6 students from School A and School B	Student questionnaire Student focus group interview

3.3 Sampling

This section introduces the sampling frame for the current study. It first describes the rationale and procedures for the sampling selection, followed by a detailed account of the participants in the current study.

3.3.1 Rationale for sample selection

The present study was set within the context of teaching English as one of the compulsory subjects to young learners from Grade 3 (starting age 9 to 10) in a sample of two state primary schools from Luoyang, Henan Province in China.

The rationale for this sampling selections was as follows. First, given the large and diverse context of China, where there are more than 200,000 state primary schools, it was neither realistic nor possible to approach a wide range of primary schools from different cities. Second, Luoyang is a prefecture-level city in western Henan Province, central China. Situated in a province with the largest population in China, Luoyang has over 113 primary schools scattered across its urban and rural areas with more than 95,000 primary school students. Unlike the first-tier cities (e.g. Beijing, Shanghai), which have introduced English as a mandatory subject since Grade 1 and where children have more contact with English outside the classroom, Luoyang is more representative of second-tier Chinese cities of middle to lower social economic status. In second-tier cities, children start to learn English from Grade 3 and have limited contact with English outside of the classroom. Although the sample in the current study is far from representative of China's population, it can nonetheless provide some in-depth insight on the teachers' curriculum implementation and assessment practices in listening and speaking and primary school students' English proficiency.

3.3.2 Procedures of sample selection

For the purposes of the present study, two state primary schools were selected based on the convenience sampling method (Cohen, Manion, & Morrison, 2007). A meeting was held with the director of the district-level Education Bureau and ELT coordinator[5] via personal contact to discuss the criteria for the sample school selection, in which 1) the schools should be the local state primary schools; 2) the schools

[5] The district-level ELT coordinator, or *Jiaoyanyuan* (教研员), is a senior official in the local Educational Bureau who is responsible for the local teacher training, curriculum implementation, assessment development and sometimes teaching material development.

should conduct the teaching under the guidance of the NECS; 3) the English course was provided since Grade 3; 4) the schools allowed the researcher to make frequent visits while research was ongoing. There are 28 state primary schools in the local district. With the help of the ELT-coordinator, the invitation letter was sent to all the local primary schools. There were 4 schools expressed their interest. After discussed about the detailed plan for the research visit with the 4 schools' principals, there were two state primary schools finally willing to participate in this study. The two schools are named as School A and School B in this thesis to maintain anonymity.

3.3.3 Participants

The target participants in this study are the main stakeholders in the classroom context. They are EFL teachers who deliver the English curriculum and students who learn English in the classroom. A detailed account of the teacher and student participants in this study are described below.

3.3.3.1 Teacher participants

Teacher participants for questionnaire

For teacher questionnaire, all the EFL teachers from the local district state primary schools were invited to participate on a voluntary basis. There were 109 EFL teachers from 28 state schools in the local district. For the main study phase, a total of 103 teachers from 28 state schools responded and completed the questionnaire. The response rate was 94.5%.

Table 3.2 presents teachers demographic information including gender, age, educational qualification, teaching experience (years), current teaching grade, total teaching hours, total class number and average student number per class.

Table 3. 2 Demographic information of teacher participants

Variables	**N**	**%**
Gender	N = 103	
Male	2	1.94%
Female	101	98.06%
Age	N= 103	
21-30	40	38.83%
31-40	47	45.63%
41-50	14	13.59%
Over 50	2	1.94%
Highest academic qualifications	N = 103	
College diploma in English	17	16.5%
College diploma in other subjects	4	3.88%
B.A. in English	60	58.25%
B.A. in other subjects	20	19.42%
M.A./M.Ed. in English	0	0%
M.A. in other subjects	2	1.94%
Teaching experience	N = 103	
1-4 years	33	32.04%
5-9 years	37	35.92%
10-14 years	14	13.59%
15-19 years	12	11.65%
Over 20 years	7	6.8%
Current teaching grades	N=142	
Grade 3	33	32.04%
Grade 4	40	38.83%
Grade 5	33	32.04%
Grade 6	36	34.95%
Teaching hours per week per teacher	N=103	
3 - 10 class hours	20	19.41%
11 - 15 class hours	77	74.75%
15 class hours and over	6	5.82%
Total class number per teacher	N=103	
1-2 classes	19	18.45%
3-4 classes	75	72.81%
5-6 classes	9	8.74%
Average students per class	N=103	
30-49 students	34	33%
50-69 students	69	66.9%

Out of all the 103 teacher participants, there were 101 were females and only 2 were males. This ratio was comparable to the findings of a survey study conducted by Wang (2009) revealing that females constituted the most part of the primary EFL teachers in China. The teachers' age ranged from 21 to 54, and their age fell predominately in two age groups: with 40 teachers (38.83%) between 21 and 30; and 47 teachers (45.63%) between 31 and 40. There were only 14 teachers (13.59%) between 41 and 50;

and two teachers (1.94%) over 51. Among the group, there were a total 84.46% teacher participants under 40 years old.

With regards to the highest academic qualifications, a majority of teachers (77.67%) held bachelor's degrees where 58.25% had degrees in English and 19.42% in other subjects. There were 20.38% teachers had college diplomas, and very few teachers (1.94%) had master's degrees. It could be noted that there is a growing trend that more teachers hold higher degrees in English in the most recent years when compared with Wang (2009)'s large-scale survey study in which there were around 20% primary EFL teachers earned a bachelor's degree with only 12% in English subject.

As for their teaching experience, most of the teachers (67.96%) had less than ten years teaching experience in which 33 teachers (32.04%) were novice. There were 26 teachers (25.24%) with 10 to 20 years of teaching experience. When asking about the current teaching grades, the total number of responses was 142 which was larger than the total number of teacher participants (103), indicating the fact that many of them were teaching with more than one grade levels.

The last three categories documented the contextual information regarding teachers' current workloads. The teaching hours required for each teacher ranged from 3 to 20 class hours per week. Of the group, most of the teachers (74.75%) taught between 11 to 15 hours a week. A majority of teachers (72.81%) taught 3 to 4 classes. There were 66.9% teachers with large class size ranging from 50 to 69 students and the remaining 33% teachers with relatively medium class size ranging from 30 to 49 students. On the whole, it revealed that most of the teachers who complete the survey reporting having mostly large class sizes and heavy workloads.

Teacher participants for classroom observation and interview

For classroom observation and interview, there were four teachers from two state primary schools were recruited[6]. One teacher at Grade 4 and one teacher at Grade 6 was recruited in each school. Background information about teacher participants is summarized in Table 3.3. The four teachers in this study were all female and all had a bachelor's degree, either in English or education. Their teaching experience ranged from 7 to 15 years.

[6] In the two participating state primary schools, at each grade level, there were usually one to two EFL teachers. In School A, there were two Grade 4 EFL teachers, however, one of them was novice teacher who only had 2 years teaching experience and thus not suitable for the current study. In Grade 6, there were only one EFL teachers available. In School B, there were only one EFL teachers in Grade 4 and Grade 6, and thus were recruited.

Table 3. 3 Background information of teacher participants

Grade level	Teacher ID	School	Gender	Age	Qualifications	Teaching experience
Grade 4	Ms. P	School A	Female	32	B.A. English	11 years
	Ms. W	School B	Female	27	B.A. English	7 years
Grade 6	Ms. L	School A	Female	30	B.A. English	8 years
	Ms. B	School B	Female	36	B.A. Education	15 years

3.3.3.2 Student participants

The student participants in this study consisted of four classes of students at Grade 4 and Grade 6 in School A and School B taught by the teacher participants. Demographic information (chronological age and gender) is summarized in Table 3.4. There were 106 fourth-grade students (Mean age=10;5, SD=0;3) and 116 sixth-grade students (Mean age=12;5, SD=0;3) in total. The gender was balanced, 50% boys and 50% girls in Grade 4, whereas in Grade 6 the percentage of boys (60%) was higher than that of girls (40%). The participating classes were similar in terms of the class size, age and gender at the same grade level.

Table 3. 4 Demographic information on student participants

		N	Chronological Age (yrs.; moths)			Gender	
			Range	Mean	SD	Boys	Girls
Grade 4	School A	54	9;5 – 11;2	10;5	0;3	29 (54%)	25 (46%)
	School B	52	9;6 -11;0	10;5	0;4	24 (46%)	28 (54%)
Grade 6	School A	56	11;7 – 12;10	12;5	0;3	34 (61%)	22 (39%)
	School B	60	11;1 – 13;2	12;6	1;3	36 (60%)	24 (40%)

It should be noted that this sample was not selected to represent the wider student and teacher population in China. Rather, this sample was chosen to provide detailed information about what students can do and what is actually happening in a sample of Chinese EFL classrooms at the primary school level.

3.4 Research instruments

This section describes the selection and adaptation of the research instruments used to collect data for this main study. Guided by the mixed-methods research design, a range of research tools were used to collect both quantitative and qualitative data.

3.4.1 Questionnaires

Questionnaires are widely used for collecting both quantitative and qualitative data from a large number of participants (Cohen et al., 2007). Efficient in terms of money and time, questionnaires are often used in social science research to investigate people's attitudes, opinions, beliefs and motivations (Creswell & Clark, 2007). In the current study, questionnaires were employed to investigate the EFL teaching and

learning experiences of teachers and students in the classroom context. The following sections describe the development of both teacher and student questionnaires.

3.4.1.1 Teacher questionnaire

This section describes the content design of the teacher questionnaire (Appendix 2) and includes a description of the pilot study carried out using the questionnaire. Further revisions and final validations are also described.

Content of the teacher questionnaire

The teacher questionnaire was used to determine the potential factors which contribute to or impede the process of curriculum implementation and to provide insights into what classroom-based assessment activities EFL primary school teachers use to evaluate their students' listening and speaking ability in Grade 4 and Grade 6. A context-appropriate questionnaire matched with this project's research focus was not available in the literature. Thus, the questionnaire items were adapted from a combination of different studies, as discussed below.

The questionnaire is divided into three sections. Section A is concerned with the demographic information of participants such as gender, age, educational qualifications, years of primary teaching experience, and their current teaching situation, including number of classes, teaching hours per week and class size.

Section B is concerned with general perceptions of curriculum implementation. It is further divided into two parts. The first part is designed to identify teachers' attitudes towards different aspects of curriculum implementation: teacher's English proficiency, teaching material, influence of exams, teaching workloads and attitudes towards the proposed teaching method. Previous studies have identified these aspects as factors that influence curriculum implementation (Bulter, 2004; Cheung & Wong, 2012; Enever, Lindgren, & Ivanov, 2014; Gorsuch, 2000; Rixon, 2013; Wang, 2006). This is composed of 15 items. Some items are adapted from Gorsuch (2000) and Wang's (2006) questionnaires measuring the impact of teaching proficiency, teaching materials, exams and teaching workloads on curriculum implementation. The items regarding teacher's attitudes towards the proposed teaching methods (i.e., communicative task-based language teaching) were adapted from Karavas-Doukas's (1996) attitude survey. Teachers were asked to rate on each statement on a 5-point Likert scale (1 = strongly disagree to 5 = strongly agree). The second part of section B asks teachers to evaluate the extent to which their current students' English proficiency is likely to reach the curriculum goals by the end of Grade 4 and Grade 6. The aim is to obtain teachers' general evaluation of their current students' performance based on the curriculum goals for listening and speaking skills. The format for this section

was adapted from Butler (2004)'s questionnaire, which was designed to investigate EFL students' English proficiency at the primary level in three East Asian countries. This questionnaire asked teachers to evaluate the importance of the English proficiency goals outlined in their respective countries' curriculum documents for 4 skills. However, instead of using the curriculum goals from those three East Asian countries, the items were adapted to reflect the NECS targets under the category of speaking and listening skills. Teachers were asked to evaluate each target on a 5-point Likert scale (1 = none of my students could achieve this target to 5 = 100% of my students could achieve this target).

Section C concerns teachers practices in assessing young EFL learners. According to the NECS, teachers are encouraged to employ multiple forms of assessment (e.g., formative assessments and summative assessments) to improve and motivate learning. This section was designed to gather information about teachers' actual assessment practices, including the assessment types, content and related activities. These items were adapted from a questionnaire investigating EFL teachers' beliefs and practices in three European countries (Brumen, Cagran, & Rixon, 2009). The original questionnaire included nine types of assessment, which address varied aspects, including content, types and sources of assessment. In the current study, these items were further divided into two parts, type of assessment and content of the assessment, which are the first and second item in Section C respectively. To determine types of assessment, the author referred to McKay's (2006) classifications of assessment for young learners including self-assessment, peer assessment, classroom observation and other types. To determine types of content, the author referred to assessment criteria recommended by the NECS, which included spelling, pronunciation, vocabulary use, grammar and other types (Chinese Ministry of Education, 2011). The original questionnaire consisted of 15 activities tapping on the listening, speaking, reading and writing. In the current study, activities related to speaking and listening were used as well as some classroom based oral related activities (e.g. listen and act) recommended by the NECS were also added.

Piloting the teacher questionnaire

The rationale for piloting the teacher questionnaire was 1) to get a sense of the procedure for data collection (e.g., time needed for completing the questionnaire); and 2) to check the construct validity by calculating the reliability coefficient. The participants in the pilot study were 48 Chinese EFL primary school teachers. The data collection was based on convenience sampling (Johnson & Christensen, 2004), a non-random sampling technique. The questionnaires were distributed at a local in-service teacher-training event in July 2016. Participating teachers were invited to complete it via an online survey tool. At the beginning of the survey, a statement was included describing the purpose of the study as well as a detailed instruction on how to fill out the questionnaire. After receiving the completed questionnaire, the responses were coded into quantifiable data and analysed by employing

SPSS 23.0 version. Descriptive statistics and Cronbach's alpha were calculated to get a sense of the data as well as to check the internal consistency of each construct. Modifications and adjustments to the final version of the questionnaire are discussed in the next section.

Final validation of the questionnaire

Based on the data analysis results of the pilot study, several adjustments and modifications were made to the questionnaire.

First, for Section B, which was designed to investigate the potential factors for the curriculum implementation, Cronbach's alpha for each factor scale was calculated to estimate the internal consistency of items under each construct. Four factors—teacher's level of English proficiency, teaching resources, influence of exams and teaching workloads—had high reliability, Cronbach's alpha was .86, .76, .81 and .78 respectively. However, one factor of the attitudes towards the proposed teaching methods had a low reliability coefficient, .3, which indicated questionable internal consistency. This factor scale included four original items designed to target on two teaching approaches, which were traditional grammar translation teaching and communicative task-based language teaching. However, consultation with local teachers indicated that the wording of two items on traditional grammatical/translation teaching were confusing. As written, they emphasized the importance of grammar in teaching speaking and listening, rather than the teaching approach itself. Therefore, the wording was changed based on their suggestions.

Second, the local ELT coordinator and some primary EFL teachers suggested local teachers' perception of NECS might also be influenced by local training, workshops and national in-service trainings. All of these could be categorised as teachers' professional developments, which is acknowledged as a potential factor of curriculum implementation in the previous literature (Enever et al., 2009; Karavas-Doukas, 1996; Khabbazbashi, Khalifa, Robinson, Ellis, & Mifsud, 2017). Therefore, four new items were added focusing on teachers' professional development.

Table 3. 5 Structure and themes of the teacher questionnaire

Structure	Themes
Section A	**Demographic information** • Gender • Age • Educational qualifications • Current teaching grade • Years of teaching experience • Teaching hours per week • Total class number • Average students per class
Section B	**B1. General perception of the curriculum implementation** • Teacher's level of English proficiency • Teaching materials (textbooks, other resources) • Influence of exams (washback effects) • Teaching time, class size and workloads • Teachers' professional development
	B2. General evaluation on curriculum targets • NECS level 1 targets in listening and speaking • NECS level 2 targets in listening and speaking
Section C	**Practices of assessment of young EFL learners** • Types of assessments • Content and skills focus in the assessments • Activities in assessing speaking and listening

3.3.4.2 Student questionnaire

This section describes the content design of the student questionnaire and includes a description of the pilot study on the questionnaire.

Content of the student questionnaire

Students views on classroom activities and how they are assessed in listening and speaking, together with their overall learning attitudes were also measured by questionnaires. There were three sections in the student questionnaire (Appendix 3). Section A concerned students' demographic information (age, gender and years of EFL learning). Section B regards students' preference about assessments activities targeted at listening and speaking skills. This section, served as a cross-checking mechanism, as it contained the same assessment activities that are specified in the teacher questionnaire, with the intention of finding out whether the students find these tasks useful. Section C is about students' general affective attitudes towards EFL learning. The items were adapted from Butler and Lee's (2010) questionnaire using a smiley face 5-point Likert-scale (e.g. 1 = strongly disagree; 5 = strongly agree). The items in this section were placed in random order to avoid leading the respondents and grouped into themes (i.e., anxiety, confidence, intrinsic motivation and instrumental motivation) in the data analysis.

Piloting the student questionnaire

The student questionnaire was piloted in one local primary school in Luoyang, China. A total of 54 Grade 4 students from one class and 53 Grade 6 students from another class participated in the pilot study. The aim of this pilot study was to 1) check whether the wording in the student questionnaire was easy to understand and 2) test the internal consistency of the affective variables in the Section C. Before the student questionnaire was distributed to the pilot group, the content of the questionnaire was first discussed with local teachers and reviewed with some Grade 4 and Grade 6 students to ensure its appropriateness to the target population in the main study (e.g., whether the content of the questionnaire was difficult to understand and whether it contained Chinese characters unfamiliar to the Grade 4 students). In terms of the internal consistency of the variables, the Cronbach's alpha of four variables in the Section C were calculated and the results suggested high internal consistency[7] of each variable at both of grade levels.

3.4.2 Interviews

Interviews are often used to acquire more in-depth information about people's perceptions, beliefs, feelings and definitions of situations (Punch & Oancea, 2014). Compared with questionnaires and observation, interviews can probe more deeply and widely, expanding on specific questions initially surveyed in a questionnaire. They can also be used to check and verify the accuracy of impressions gained through observations (Fraenkel & Wallen, 2007). Therefore, interviews were used to provide more comprehensive insights into teachers' views on teaching and assessing oral and aural skills and students' experience in EFL learning.

Interviews can be classified into three categories: the structured interview, the semi-structured interview, and the unstructured interview (Merriam, 1998). Structured interviews are verbal questionnaires to test a hypothesis or to survey a sample of particular respondents. The questions are usually fixed and are formatted in a predetermined order (Fraenkel & Wallen, 2007; Punch & Oancea, 2014). Semi-structured interviews are guided by the research questions in a more flexible manner, where the order of the questions and the way to probe the issues is not determined beforehand (Merriam, 1998). Unstructured interviews, also called ethnographic interviews, are open-ended and non-standardised, and the process is directed by the respondent (Punch & Oancea, 2014). In the current study, semi-structured interviews were used to guarantee that main research questions are answered while allowing the respondents to

[7] In the pilot study, Cronbach's alpha for four variables in Grade 4 was as follows: .78 for anxiety, .76 for confidence, .71 for intrinsic motivation and .66 for instrumental motivation. The Cronbach's alpha for four variables in Grade 6 was as follows: .76 for anxiety, .88 for confidence, .81 for intrinsic motivation and .70 for instrumental motivation.

express their thoughts freely and the researcher to probe into areas that are worth investigating. The interview data was also seen as supplemental information used to triangulate with data obtained from questionnaires and classroom observations.

3.4.2.1 Teacher interviews

Semi-structured interviews were conducted with each of the observed teachers individually. The interview schedules (see Appendix 4 for the list of teacher interview questions) included four sections: Section 1 collected relevant background information about the teacher. Section 2 asked about teachers' general understanding and views of the NECS and the potential factors which might influence their implementation. Section 3 focused mainly on teachers' daily teaching practices, especially as they taught listening and speaking. Section 4 investigated teachers' understandings and perceptions of the multiple assessments recommended by the NECS, and their actual enactment on multiple assessments.

Four teachers from two sampled primary schools were interviewed in the main study, two teachers from Grade 4 and two teachers from Grade 6. The interview took between 40 and 50 minutes. The information sheets as well as the consent forms were distributed to teachers before interview. The interview was conducted in Chinese and audio recorded with the consent of participants.

3.4.2.2 Student interviews

A series of focus groups were conducted with students to gather data on student perspectives of classroom-based assessment activities. Student viewpoints are important because they are the ones who experience the classroom activities, and the interview allowed students' voices to be heard (Enever, 2011). The interview schedules (Appendix 5) with students included 11 open-ended questions addressing students' views on classroom activities, students' views and preference on classroom-based assessments and students' views on teacher feedback.

The criteria for student selection were as follows. First, the students needed to be expressive to ensure a minimum amount of data. Second, they needed to represent different levels of EFL learning ability and different genders to represent diverse perspectives on the learning and assessment activities. Third, for each focus group, there were six students involved. The number of each group was within the recommended range of six to eight (Barbour, 2008). The teachers were consulted when selecting the student participants given that they had taught these students since Grade 3 (or nearly one academic year) and they were more familiar with their students' general performance and language ability. There were six low achievers and six high achievers from each class. Students at both levels were identified by their teachers, and these assessments were later confirmed by the researcher during classroom observation and individual assessments.

Approval for students' participation in the interview was obtained from the school principal, their teachers and their parents. The interviews were conducted in Chinese and audio recorded with the participants' consent. Interview data was transcribed and translated for data analysis. Each student from the focus groups was assigned a unique identification code to preserve confidentiality. For example, H1_G4SA indicates a student from Grade 4 School A as a high achiever, whereas L1_SBG6 indicates a student from Grade 6 School B as a low achiever.

3.4.3 Classroom observations

Classroom observation, as one of the essential data collection techniques for the mixed-methods research, provide first-hand information in the actual research site (Creswell & Clark, 2007). In addition, classroom observation would facilitate better understanding of the context. It allows the researcher to crosscheck issues or notice things that might be missed during the survey or interview (Merriam, 1998). Therefore, a series of classroom observations were carried out in the current study. The use and adaptation of the classroom observation instrument are described below.

3.4.3.1 Identifying the classroom-based aural and oral assessment activities

In order to identify the classroom-based aural and oral assessment activities, a discourse analysis on the classroom transcripts was conducted to see if the classroom activities involved of three core steps: eliciting student's learning evidence, interpreting student's response, and providing feedback to student's response. These three core steps have been identified in the previous literature (Black & Wiliam, 2009; Davison & Leung, 2009c; Hill & McNamara, 2011). As being key elements in the implementation process of classroom-based assessment, they were usually found in the format of Initiation-Reply-Evaluation or IRE (Mehan, 1979). In the first step, eliciting student's learning evidence, the teacher might involve students to participate in a certain activity or task by informing them about the task aims or clarifying the criteria for success (Black & Wiliam, 2009; Hill & McNamara, 2011). This step usually begins with clear discourse markers, for example, "Now" and "OK" to indicate the beginning of an assessment activity (Wang, 2017). The second step is interpreting the student's response. During an assessment activity, either the teacher or students might role as the agent of judgement, focusing and evaluating on students' actual performance in relation to prespecified activity objectives (Hill & McNamara, 2011). The third step is providing feedback to student's response. Teachers make use of their judgement by giving feedback to individuals, groups, or the whole class with the actual content of feedback at either "person-referenced" or "task-referenced" level (Black & Wiliam, 2009). For example, the following two extracts, taken from one observed Grade 4 lesson by Ms P, represented a classroom-based assessment activity and a non-assessment one respectively. In Extract 1, Ms P asked her students to do a listen and answer activity. Before she started, she specified its aims and elicited student's learning evidence by asking them to watch a video with the pre-set question in mind (the first step). After that,

she focused on and interpreted student's actual response (the second step), and then provide feedback accordingly (the third step). Since all three core steps were clearly identifiable, this activity was thus regarded as a classroom-based assessment one. In contrast in Extract 2, during a listen and follow activity, although the teacher specified the task aims and elicited student's learning evidence by asking them to read after the dialogue while listening, there were no follow up steps identified regarding the teacher's interpretation and feedback on students' performance. Instead, she moved to the next activity. Therefore, this activity was not regarded as an assessment one. In the current study, a series of classroom-based assessment activities were identified based on these three key steps.

Extract 1 (Ms P, Grade 4) Episode of classroom-based assessment activity

T:	Now, please have a look at the video, and listen to what they say. After that, I will ask you what you see. Ok? Let's go!
	[Teacher played the video to the whole class]
T:	Ok, boys and girls, so many school things, right? So, what did you see?
S:	I see some colour chalks.
T:	Good, I SAW some colour chalks. Do you have some colour chalks?
S:	No.
T:	No? you....
S:	No, I don't.
T:	OK, good. Sit down please. What else, what did you see? Anyone?

Extract 2 (Ms P, Grade 4) Episode of non-classroom-based assessment activity

T:	OK, boys and girls. Now, open your book and turn to page 45. Let's listen and follow the recordings.
Ss	[Students followed the recordings and read after the dialogue in the textbook]
T:	Ok, next. Look at the blackboard. Let's practice these sentences in your group one by one, OK?

3.4.3.2 The adapted COLT Scheme

After having identified the classroom-based aural and oral assessment activities, a classroom observation scheme was used to find out 1) what were the features of the identified classroom-based assessment activities and 2) how teachers implemented these activities in related with the NECS targets. The Communicative Orientation of Language Teaching (COLT) observation scheme was adapted for the current study (Spada & Fröhlich, 1995). This observation scheme was developed within a context of a project investigating the nature of second language proficiency and its development in classrooms. It served as an instrument to systematically describe different aspects of instructional practices and procedures in classrooms. The COLT observation scheme was suitable for the current study because it helped with documenting assessment activities and analysing the associated features of these activities happened in the classroom context.

The COLT observation scheme was divided into two parts. Part A described classroom events at the level of episode and activity. Part B analysed the communicative features of verbal exchanges between

teachers and students and/or students and students as they occurred within each episode or activity. However, the current research aimed to look at the actual assessment practice at the activity level and not focus on examining the communicative features of classroom interaction. Thus, Part A was adapted accordingly.

Part A differentiated seven main features: time; activities and episodes; participant organisation; content; content control; student modality; and material (Spada & Fröhlich, 1995). The time and activities features were used to collect various types of activities/tasks. The remaining features were used to analyse the specific activity/task into details from different perspectives. Each of these features contains a number of sub-categories.

The adapted COLT scheme consisted nine categories (Appendix 6). The first two categories were time and activities, which was adopted directly from original COLT scheme. The remaining seven categories were divided into two parts, which were described in detail in the following two sections. For the first part, there were four categories were adapted to describe the features of classroom-based assessment activities. This was corresponding to the first step of a classroom assessment activity: what activity was used to elicit student's learning evidence. For the second part, three categories were added to describe the implementation practices of these assessment activities. This was matched with the second and third steps: how student's evidence was interpreted and what judgement/feedback was made accordingly.

Describing the classroom-based assessment activity

Four categories were modified from the COLT scheme: *Participant organisation*, *Content*, *Student modality* and *Materials* used. These four aspects were used to describe the characteristics of assessment activities in listening and speaking.

Participant organisation was used to distinguish whether an aural and oral assessment activity was teacher-centred or learner-centred. Group work was considered to be essential in helping students developing their communicative competence. Students had the opportunities to co-operate each other to engage in genuine interactions using a greater variety of linguistic forms and to develop their oral fluency skills (Spada & Fröhlich, 1995).

The category *Content* was used to document whether the primary focus of an aural and oral assessment activity was on forms, meanings or in combination. Traditional EFL/ESL instruction emphasised on vocabulary memorization and grammar correction. However, this was not viewed as a successful approach when developing students linguistics and communicative competence (Spada & Fröhlich, 1995). Using this category, it could be used to analyse what was the focus of an assessment activity. The original *Content* category had four sub-areas: *Management*, *Language*, *Topic* and *Content control*.

However, during the pilot study, three sub-areas *Management*, *Topic* and *Content control* were omitted. The reasons were as followed. First, *Management* was not the primary focus of current research. Second, all the teachers in the pilot study based their teaching on the textbook. The range of *Topic* was narrow and related to students' daily life, which need not to be examined. Third, since the students were very young, it was noted that they could not exert any *Content control* over topics nor suggest topics for future teaching. The *language* area was simplified into three sub-categories (i.e. form, meaning, combination) to suits the needs of current study.

Student Modality was used to identify which particular skill area or combination of skills was involved in a classroom assessment activity. Since the focus of current study was on listening and speaking skills. Therefore, the sub-categories were reduced into three: listening, speaking and combination.

The category *Materials* was used to describe the types and sources of classroom-based assessment activity materials. One sub-category *Type of materials* was kept from the original COLT scheme and were classified as written text (minimal and extended), audio and visual. However, during the pilot study, it was noted that the textbook currently used by the EFL primary students was officially chosen and approved by the government, the other sub-category *Source of Material* did not need to be examined.

Describing the implementation practices of the classroom-based assessment activity

Three categories were added into the adapted COLT scheme including *Agent*, *Target* and *Actual Focus of Assessment*. These categories were the practical aspects mentioned in both Black and Wiliam's (2009) and Hill and McNamara's (2012) frameworks describing how students' learning evidence was assessed.

Agent of Assessment was used to describe who was conducting the assessment. Two sub-categories were included to distinguish whether teacher acted as the sole agent of assessment or peers were also involved during the course of assessment. As has been discussed in the literature review, peer assessment might have the potential benefits for learners act as the active ones to monitor and evaluate other's learning performance and enhance their language competence (Black & Wiliam, 2009; Topping, 2009).

Target of assessment was used to document whose learning evidence was assessed. It comprised three sub-categories representing the response made by the whole class, by a group of students, or by individual students. When whole class or group of students' choral response were elicited as the assessment target, students' actual language competence might be disguised. It revealed limited information about individual students' learning outcomes (Holliday, 1994). Student individual responses, however, might facilitate effective assessment to check student's learning status, including problems or progress (Yan, 2014).

During the classroom-based assessment, feedback was viewed as a bridge to connect where the learner was right now and how to get there (Black & Wiliam, 2009). The content of feedback reflected the evaluator's actual focus of students learning outcomes. The category *Actual focus of assessment* was used to distinguish to extent to which the feedback focused on the use of linguistic forms, or on the meaning expressed or on a combination of the two.

In sum, this component of observation is essential as the adapted COLT provides a picture of what activities the participating teachers use for assessing listening and speaking how they actually conduct their teaching in classrooms.

3.4.4 Assessments

In this section, the use of assessments was intended to address the fourth main research question, *what can young EFL learners actually do in speaking and listening upon finishing Grade 4 and Grade 6 according to the targets set by the NECS?* The first stage involved identifying the corresponding skills of the NECS targets and trying to find suitable assessment instruments targeting those skills. The details of the assessment instruments were discussed and followed with a description of the pilot study. The second stage focused on describing the process through which a panel of local teachers linked the assessment results with the NECS targets. The methods used in the linking process are justified and discussed.

3.4.4.1 Assessment instruments

Two assessment instruments were employed to assess the actual oral and aural learning outcomes of Chinese EFL learners at the primary school level. The NECS stipulates that Grade 4 and Grade 6 students are expected to achieve a series of targets, which have been detailed under the domain of four skills. The targets in the domain of listening and speaking are those of two externally developed assessment instruments deemed suitable for use with both Grade 4 and Grade 6 students: the British Ability Scale-II (BAS II; Elliot, Smith, & McCulloch, 1996) and the storytelling task *Frog where are you* (Mayer, 1969).

Table 3.6 illustrates the targets in Level 1 and their corresponding knowledge and skills, as well as the related assessments used to assess students' listening and speaking skills. The targets in Level 2 and their associated assessments are listed in Table 3.7. The assessments were assumed to tap into the constructs that are meant to be identifiable in the can-do statements. The description of each instrument and the reasons for use in the current study were detailed in the following sections. Due to the difficulty of finding standardised or widely used assessment tools for assessing pronunciation, functional knowledge and communicative skills, this study focuses on aural comprehension and oral production (i.e., narrative skills)

Table 3. 6 NECS targets in Level 1, contributing knowledge and skills, and research instruments

Skills	Targets	Contribution knowledge & skills	Research instruments
Listening	1. Can recognise pictures or real objects based on heard vocabulary and sentences	Understanding vocabulary	BAS-II Verbal Comprehension: Section 1 & 2
	2. Can understand and react to simple classroom instruction. 3. Can take action based on directions. (e.g., pointing at pictures, colouring, drawing and acting).	Understanding instructional language	BAS-II Verbal Comprehension: Section 2
	4. Can understand simple stories with clues from pictures and actions.	Listening comprehension at discourse level	BAS-II Verbal Comprehension: Section 2
Speaking	1. Can imitate speaking based on the tapes.	Pronunciation	N/A
	2. Can greet in simple sentences. 3. Can exchange personal information (e.g., name and age). 4. Can express simple emotions (e.g., like and dislike). 5. Can guess the meaning of the words based on acting.	Functional knowledge, communicative skills	N/A
	6. Can produce words, phrases and simple sentences based on pictures.	Narrative skills	Frog storytelling task

Table 3. 7 NECS targets in Level 2, contributing knowledge and skills, and research instruments

Skills	Targets	Contributing knowledge & skills	Research instruments
Listening	1. Can listen and understand conversations and recordings with clues from pictures, images or gestures.	Listening comprehension (dialogue)	BAS-II Verbal Comprehension: Section 3
	2. Can listen and understand picture stories.	Listening comprehension (text)	BAS-II Verbal Comprehension: Section 3
	3. Can understand simple questions in class. 4. Can understand common instructions and directions and response properly.	Understanding instructional language	BAS-II Verbal Comprehension: Section 1 & Section 2
Speaking	1. Can articulate English clearly with the appropriate intonation.	Pronunciation	N/A
	2. Can have simple conversations based on familiar personal or family situations. 3. Can use daily expressions (e.g., greetings, express one's thanks, apologies) and express simple emotions (e.g., like and dislike).	Functional knowledge, communicative skills	N/A
	4. Can narrate stories based on daily topics. 5. Can narrate simple stories with the help of teachers and clues from pictures.	Narrative skills	Frog storytelling task

Listening test: British Ability Scale-II Verbal Comprehension

The British Ability Scales (BAS II; Elliot, Smith, & McCulloch, 1996) was used as an instrument to assess Chinese EFL children's listening comprehension for the following reasons. First, with the exception of the listening part of large-scale tests (e.g., the Cambridge YLE tests[8]), there are no specific standardised tests available to test listening comprehension in EFL learners, let alone EFL children in China. Second, the BAS II has been used to assess EAL and EFL children's language skills in previous research (Dockrell, Stuart, & King, 2010; Hutchinson, Whiteley, Smith, & Connors, 2003; Sen & Blatchford, 2001) and is equally applicable to this study's sample population. Third, the BAS-II verbal comprehension subtest measuring receptive language through understanding of oral instructions uses basic language concepts that match the Level 1 and 2 descriptors in the listening skills category from NECS. According to NECS, Chinese EFL children are expected to understand common instructions or directions and respond properly (Chinese Ministry of Education, 2011).

The BAS-II Verbal Comprehension contains 40 items in total (Appendix 7). This includes four sets of items (i.e., teddy bear exercise; inset tray exercise; shapes exercise and listening comprehension exercise) that mainly focus on assessing children's ability to understand instructions and comprehend sentences. Students were asked to manipulate objects or point to pictures in response to instructions from the administrator. For each student, the total score represented the sum of correct response scores on a particular range of items administered. There were 23 items administered to Grade 4 students and 33 items administered to Grade 6 students, yielding a maximum score of 23 and 33 respectively. Although the relative standard/scale scores were determined by the BAS-II, those scores were normed in the U.K for use on monolingual children between 3 to 7 years old. However, they were not meaningful for the current participants. The details of the pilot study and the modified version of the BAS-II Verbal Comprehension for the main study phase are discussed in piloting results section.

Speaking test: Frog storytelling task

A narrative task was used as an instrument for assessing Chinese EFL primary students' oral language skills for the following reasons. First, narrative tasks have been used in numerous studies to assess oral language competence, since they enable assessment of students' English as they use it at the discourse level (Cowley & Glasgow, 1994; Lasagabaster, 2011; Petersen, Gillam, & Gillam, 2008). Rich information, including vocabulary and syntax during the narrative task, are gleaned using this method.

[8] The Cambridge YLE tests were excluded from the current study for the following reasons: 1) Although the tests were designed for young EFL learners around the world, they are based on an external curriculum and 2) they have their own rationale and scoring rubrics for assessing specific skills at different levels.

Since a narrative task includes a prompt that requires students to think, retrieve their vocabulary and use their language, it can also be seen as a communicative event. Second, narrative skills are one of the targets specified in the NECS under the domain of speaking skills, and students are expected to be able to narrate simple stories based on the pictures they are shown. Grade 4 students are expected to narrate simple phrases at the sentence level. Grade 6 students are expected to narrate at the text level (Chinese Ministry of Education, 2011).

Therefore, the current study used a 24-page wordless picture book, *Frog, where are you?* (Mayer, 1969) for the storytelling task. This book is about a boy and his dog who are looking for their missing frog (Appendix 8). Since the book contains no words and consists of rich but easily understood episodes, this task is commonly used to elicit narratives from children of different language backgrounds(Kang, 2012; Lasagabaster, 2011; Nicolopoulou, 2008). In addition, since every participant narrates based on the same content, the picture book elicitation enables relatively reliable comparison across participants. During the task, both Grade 4 and Grade 6 students were asked to tell a story based on this picture book. Their narratives were audio recorded.

Participants' narrative samples were transcribed verbatim in CHAT format and analysed using the Computerised Language Analysis programme (CLAN) (MacWhinney, 2000). Narrative skills were analysed at the macrostructure and microstructure levels (Westerveld & Gillon, 2008). Macrostructure refers to overall quality of the narratives that students produced, including story grammar elements that contribute to the general framework of the narratives (Stein & Glenn, 1979), whereas microstructure refers to the linguistics level, such as the fluency (e.g., total number of words), length (e.g., mean length of utterance) and complexity (e.g., total number of different words) aspects (Petersen, Gillam, Spencer, & Gillam, 2010). The Index of Narrative Complexity scoring scheme (Appendix 9) (Petersen, Gillam, & Gillam, 2008) was used to analyse and score the macrostructure of participants' narratives as this tool (1) was primarily designed based on Mayer (1975)'s frog story series (e.g. *Frog where are you*; *One frog too many*); (2) integrated both story grammar and linguistic related elements of oral narratives and (3) was designed to measure changes in the narratives (Petersen et al., 2008). Following the index, narratives were rated based on a 3-point scale (scores ranging from 0 to 2 or 0 to 3) targeting 13 narrative elements: character; setting; initiation event; internal response; action and consequence; plan; complication; consequence; formulaic markers; temporal markers; casual adverbial clauses, knowledge of dialogue and narrators' evaluations. Scores for each of these 13 elements were added up to form a final score of 31. To establish the inter-rater reliability in the main study phase, approximately 20% of Grade 4 (n = 22) and Grade 6 (n = 21) students' samples were randomly selected and rated independently by a PhD student in Education who is a native speaker of English. The inter-rater agreement for Grade 4 and Grade 6 students' oral samples were 0.92 and 0.94, respectively. In addition, the microstructure

of participants' narratives was analysed in terms of mean length of utterance (MLU), total number of words (TNW) and total number of different words (TNDW).

Piloting the assessment instruments

A pilot study of the assessment instruments was conducted from early June 2015 until mid-July 2015 (3 weeks in total). The aims of the pilot study were (1) to help the researcher decide on selected instruments, (2) become familiar with the testing procedures, (3) estimate the testing time required for each participant and (4) conduct a preliminary analysis of the pilot data. The details of the pilot study are presented below.

Participants in the pilot fieldwork were Grade 4 and Grade 6 students from one local primary school in Luoyang City, Henan Province, China who were recruited using a convenience and snowball sampling method (Cohen et al., 2007). A total of 29 Grade 4 students (Mean age=10;5, SD=0;6) and 30 Grade 6 students (Mean age=12;5, SD=0;4) were tested individually. The students were drawn from two classes, one Grade 4 class and one Grade 6 class. This pilot sample was comparable to the main study sample in terms of demographic characteristics and experience with EFL learning in a school context. Participants who took part in the pilot study did not join the main study.

The pilot fieldwork was scheduled near the end of the academic year 2014/2015 to allow Grade 4 participants to complete a full 2 years of EFL learning and Grade 6 participants, a full 4 years. In order to accommodate various school calendars and timetables, data collection was conducted on Grade 6 first, followed by Grade 4. Data collection consisted of two individual assessments: the BAS-II verbal comprehension and the storytelling task. The testing setting was arranged to ensure children were at ease. A certificate of participation was given out to children who completed all sessions. Descriptive statistics were computed for each measure. The results of Grade 4 and Grade 6 are summarized in Table 3.8 and Table 3.9 separately. No missing data was reported in the pilot study.

Table 3. 8 Descriptive statistics for Grade 4 students in pilot study

Measures	**No.**	**Range**	**Mean**	**SD**	**Skew.**	**Kurt.**
BAS-II						
Verbal comprehension	29	5-33	11.28	5.53	2.28	7.71
Storytelling Task						
Macrostructure						
INC score	17	4-16	8.71	4.22	0.89	-0.77
Microstructure						
MLU	17	4.52-6.2	5.56	0.63	0.09	-1.89
TNW	17	88-198	127.23	34.97	1.04	0.10
TNDW	17	27-73	41.58	13.41	1.14	0.48

Note: INC=Index of Narrative Complexity; MLU=Mean Length of Utterance; TNW=Total Number of Words; TNDW=Total Number of Different Words

Table 3. 9 Descriptive statistics for Grade 6 students in pilot study

Measures	No.	Range	Mean	SD	Skew.	Kurt.
BAS-II						
Verbal comprehension	30	3-28	12.83	6.77	0.65	-0.23
Storytelling Task						
Macrostructure						
INC score	17	5-19	12.88	4.45	-0.39	-1.19
Microstructure						
MLU	17	5.18-7.84	6.319	0.81	-0.54	-0.38
TNW	17	61-298	196.11	70.66	-0.45	-0.87
TNDW	17	30-107	68.70	21	0.133	-0.57

Note: INC=Index of Narrative Complexity; MLU=Mean Length of Utterance; TNW=Total Number of Words; TNDW=Total Number of Different Words

As the primary purpose of this pilot study was to test research measures, distributions of scores were examined for normality. Descriptive statistics for all variables (including range, mean, standard deviation, skewness and kurtosis statistics) were run to better understand raw data from the pilot study (Field, 2009). Skewness is a measure of the symmetry of a frequency distribution. Positive values of skewness indicate too many low scores in the distribution, reflecting that the measure was too hard for the children. Negative skewness indicates a higher number of high scores earned, indicating a relatively easier measure for the children. Kurtosis measures whether a distribution is heavily tailed or light tailed, which indicates the number of outliers. Statistical assumptions were checked to see if skewness and kurtosis values fell between 2.00 and –2.00.

As shown above in Table 3.8 and Table 3.9, for the BAS-II verbal comprehension, Grade 6 students' scores were normally distributed, but Grade 4 students' scores were slightly skewed. This might indicate that this assessment instrument might be difficult for the Grade 4 students. For the storytelling task, there was no evidence of skewness and kurtosis found in either Grade 4 and Grade 6 students' scores at the macrostructure or microstructure level. Although results from the pilot study have limited scope due to the small numbers of students involved in both tests, the study provides insight into the suitability of the selected instruments for the population targeted in the main study.

Several modifications were made based on the procedures and findings from the pilot study. First, the coloured shape items in the BAS-II Verbal comprehension were excluded. Most of the students were not able to attempt the shapes exercise due to their lack of familiarity with shape-related vocabularies. Consultation with local EFL teachers revealed that key words relating to shapes (e.g., triangle and square) were neither included in the wordlist of the participating schools' textbooks nor in the NECS. This might explain that the skewed scores obtained from Grade 4 students during the pilot study. Second, during the storytelling task, some students found it too challenging to tell a story spontaneously. Therefore, the main study was designed to allow 5 minutes for the students to go through all the pictures and discern the structure of the story before the narration task. Additionally, Grade 4 children were

found to have short attention spans, easily showing signs of fatigue. The main study phase was designed to arrange students into groups where 2 to 3 children were tested in turn, with each session not lasting more than 10 minutes. In contrast, Grade 6 children, with longer attention spans, could complete all the testing in a longer session of approximately 15-20 minutes.

3.4.4.2 Standard setting by an expert panel

This section describes the approach and procedures followed in this study to interpret the two individual assessment results in relation with the NECS by soliciting expert input from a group of local teachers.

Standard setting

The linking process followed in this study to map the two individual assessment scores onto the targets set by the NECS is called *standard setting*. Standard setting includes a number of approaches (e.g. the bookmark method, the Angoff method) that are widely used to establish "cut scores" (i.e. cut-off scores) for tests based on judgements by a panel of experts (Cizek & Bunch, 2007, p. 15). These experts, who are familiar with the characteristics of the examinee population and the content of the tests, determined whether these examinees have reached the criteria of a specific performance level. Since the cut-off scores are derived from the opinions of the judges, the standard setting is not aimed at revealing some pre-existing or "true" cut-off score. Rather, it represents a translation process where educational standards or criteria are translated into the numerical language of the test score (Kane, 2001). In the current study, the standard setting procedures were followed to identify cut-off scores for two individual assessments, the BAS-II Verbal Comprehension and Frog storytelling task, corresponding to Level 1 and Level 2 of the NECS.

Regarding the general procedures of implementing standard-setting methods, this study complied with the preliminary guidelines in The Manual for Relating Language Examinations for the Common European Framework of References (CEFR) for Languages (hereafter The Manual) (Council of Europe, 2009).The Manual described four stages of the procedure: *familiarisation* with the CEFR, *specification* of the examinations in relations to the levels of CEFR, *standardisation* of the expert judgments and *empirical validation* (Council of Europe, 2009). This linking study emphasised *familiarisation* and *standardisation* of the judgments. In this phase of the study, a panel of local teachers identified the cut-off scores of two individual assessments and used the cut-off scores to differentiate a local sample of Chinese primary EFL students according to their levels in the NECS. Given this aim, linking results was not considered to be generalizable to the whole population of primary EFL learners in China. Rather, it was seen as a means of developing a richer picture of local context. Additionally, since the two individual assessments used in this study were developed externally and not designed to map onto all levels of the NECS, this study does not explore whether the content of the two individual assessments

matches with NECS descriptors. Nonetheless, the two individual assessments were designed to independently assess specific aspects of English ability of interest to the researcher (i.e., listening and speaking skills) and the overlap between the skills they tested, and the proficiency targets described in the NECS was sufficient for the linking purpose.

Selection of panelists

Participants in the standard setting process are called *panelists* (Cizek & Bunch, 2007). Panelists are the judges who make estimations and identify the cut-off scores. In the current study, it was important that the panelists 1) were familiar with the test content; 2) had a good understanding of the NECS; 3) had rich experience teaching primary-school Chinese EFL learners at different grade levels. Additionally, as recommend by Cizek and Bunch (2007, p. 107), the panel should include at least one panel member who has "personal knowledge of real examinees' levels of knowledge or skill level with respect to the characteristic being assessed". In this study, 4 teachers of the participating students were invited to be part of the panel. They also met the three aforementioned criteria for panelists. Since the recommended number of panelists is between 10 to 15 (Hurtz & Hertz, 1999), additional 8 local teachers from other state primary schools in the local district were also selected as panelists for the current study. There were altogether 12 local teachers were recruited for this aspect of the study. Each panelist was asked to complete a background information form on their qualifications and teaching experience.

Table 3.10 presents the demographic characteristics of the 12 teachers. Panelists' ages ranged from 27 to 41 and they had teaching experience ranging from 7 to 18 years. They all had obtained a bachelor's degree either in English or in Education with the associated English proficiency qualifications. They were recruited from 6 state primary schools in the local district.

Table 3. 10 Demographic information on panelists

Panelist ID	School	Age	Gender	Years of teaching	Educational qualifications	Job title	English proficiency qualifications
P1	A	32	Female	11	B.A. English	H	TEM-4
P2	A	30	Female	8	B.A. English	E	TEM-4
P3	B	27	Female	7	B.A. English	E	TEM-4
P4	B	36	Female	15	B.A. Education	H	CET-6
P5	C	41	Female	18	B.A. English	E	TEM-4
P6	C	29	Female	9	B.A. English	H	TEM-4
P7	D	39	Female	17	B.A. English	E	TEM-4
P8	D	32	Female	10	B.A. English	E	TEM-4
P9	E	28	Female	8	B.A. English	E	TEM-4
P10	E	35	Female	14	B.A. Education	H	CET-6
P11	F	31	Female	8	B.A. English	E	TEM-4
P12	F	38	Female	16	B.A. Education	E	CET-6

Note. H stands for Head of English department[9]; E stands for English teacher; TEM-4 stands for Test for English Major band 4[10]; CET-6 stands for College English Test band 6[11].

The approach used to identify the cut-off scores of two individual assessments is described in the next section.

Procedures for identifying cut-off scores

Two standard-setting methods were used to establish cut-off scores for the BAS-II Verbal Comprehension and the frog storytelling task: the modified Angoff method and the contrasting group method. The reasons for choosing these methods and the procedures of implementing them in the current study are discussed in this section.

[9] *Head of English department* refers to the English teacher who serves as the subject leader in their local primary school. The department head organises weekly or monthly meetings with other English teachers in the school.

[10] English major students must pass the TEM-4 in order to receive the degree.

[11] Non-English major students are required to pass the CET-4 for graduate and are encouraged to pass the CET-6 to prove higher level of English proficiency and for better job opportunities. CET-6 is the highest level of English proficiency test for non-English majors.

The modified Angoff method

The BAS-II Verbal Comprehension test targets receptive skills (i.e., listening skills), and the total score of each student equals the total number of overall correct responses. The modified Angoff method[12] was adopted for the following reasons. First, the Angoff method is one of the most commonly used standard-setting methods and uses a test-centred approach where panel members make judgements based on the test-item itself. Second, this method is suited for use with multiple-choice tests or tests with items (Cizek & Bunch, 2007). The BAS-II Verbal Comprehension contains the items that requires students to manipulate objects or point to pictures in response to oral instructions. This test format is similar to a typical listening comprehension test, in which students have several choices available (e.g., a student is faced with several objects and they need to pick up the correct one based on the oral instruction). The modified Angoff method usually asks a panel of experts to estimate the probability that a "minimally acceptable person" at a certain level could answer the item correctly (Cizek & Bunch, 2007, p. 83). In this study, a panel of local teachers was asked to estimate the probability of a just-qualified candidate at Level 1 or Level 2 in NECS to answer the test item correctly. The just-qualified candidate represents a student who has just enough skills to meet the targets set by the NECS at a certain level.

The following stages were carried out during a series of meetings led by the researcher.

Stage 1: Familiarisation

The familiarisation stage aims to give the panelists a clear understanding of the study's purpose and of the level descriptors in the NECS. This stage followed the steps outlined in a study by Tannenbaum and Wylie (2008): First, before the panel met, the panelists were asked to review the level descriptors of listening skills in the NECS and to write down the knowledge and skills possessed by a just-qualified EFL primary student at each NECS level, drawing on their teaching experience and knowledge of the context where the test was administered. Second, during the meeting, the panelists were asked to share their pre-meeting summary of the just-qualified student and to develop more detailed and specific descriptions of the listening skills required for each level based on the whole panel's discussion. A set of agreed-upon criteria for listening skills was established for 2 levels: Level 1 and Level 2. These definitions served as the reference for the later standard-setting judgement. In addition, the panel

[12] The modified Angoff method replaces the original Angoff method, first introduced in 1971 by William Angoff in a footnote of a book chapter. It asked the expert panel to estimate the probability that a "minimally acceptable person" could answer each item correctly. Each item was scored 1 (correctly) or 0 (incorrectly) and the sum of raw scores earned by this "minimally acceptable person" was treated as the cut score (Angoff, 1971).

members were introduced and acquainted with the BAS-II Verbal Comprehension test. They were encouraged to try the test by themselves to have an overview of it in its entirety.

The familiarisation results regarding the agreed-upon definitions are reported in the research finding chapter.

Stage 2: Standard-setting judgement

During the standard-setting stage, the panelists were first trained to practice their judgements before the actual rating. They were then asked to review each test item individually and to estimate the probability that a just-qualified student at Level 1 and Level 2 would answer the item correctly. They were reminded to consider the agreed-upon definitions while rating. To clarify what was meant by *probability*, the panelists were asked to consider 100 just-qualified students at each level and estimate how many of them (i.e., out of 100) would answer the test item correctly. They made the judgements using the percentage with increments of 5 (e.g. 35% or 70%) and marked their estimate on a rating sheet. The higher the estimated probability, the easier an item should be for the just-qualified student at a given level. The probabilities over items for each panelist were summed to arrive at a proposed cut-off score. The panelists' individual cut-off scores were then averaged to arrive at a recommended cut-off score for the whole panel.

To ensure the validity of evidence obtained during the standard setting process, two rounds of judgements were included. Each round was followed by a panel discussion on the cut-off score recommendations. Before second round, the standard deviation among the panelists for each item was calculated by the researcher and the actual percentage in answering each item correctly was provided. A common principal applicable to the modified Angoff method is that a low standard deviation indicates agreement among the panelists and a trustworthy proposed cut-off score (Hambleton & Pitoniak, 2006). The actual probability that a just-qualified student would answer each item correctly reflects the item difficulty for the current population. During the second round, the panelists were invited to discuss those items with a high standard deviation and possible reasons for the discrepancies in estimation. They were then asked to consider making changes to their initial ratings. The cut-off scores were finalised based on the second-round judgements.

The contrasting group method

To link the results of the frog storytelling test to the NECS, the contrasting group method was applied. In contrast with the modified Angoff method, in which the judgements were made primarily based on the test items, the contrasting group method is appropriate for tests targeting productive skills (e.g., speaking or writing) that judge the examinee's actual performance (Livingston & Zieky, 1982). In the

contrasting group method, the panelists, without knowing the examinee's actual test score, assess the examinee's performance by labelling it "mastery" or "non-mastery" at a given level (Cizek & Bunch, 2007, p.107). The cut-off score for each level was derived by comparing the distributions of the test scores for the examinees who were classified into each category (mastery/non-mastery) (Cizek & Bunch, 2007). For the participating panelists, judging students' oral production based on their teaching experience was a familiar task. Moreover, the students in the sample were well known to some panel members and belonged to a similar population to the students taught by the rest of the panelists. This familiarity with the students and context is a requirement for this method to ensure more meaningful outcomes.

The following stages were carried out during a series of meetings led by the researcher.

Stage 1: Familiarisation

As with the previous familiarisation stage, during the standard-setting process for the listening levels, the panelists were asked to (1) review the NECS descriptors of speaking skills at Level 1 and Level 2 and write the key characteristics of the speaking skills of a just-qualified student in at each NECS level based on their teaching experience; (2) share their personal understandings during the meeting, and discuss the speaking skills required to designate someone a qualified candidate with the other panelists. They were also invited to develop more specific descriptors of the speaking skills at Level 1 and Level 2 based on their local context. The agreed-upon definitions of speaking skills are reported in the research findings chapter. Additionally, panelists were asked to familiarise themselves with the frog storytelling task, and they were encouraged to try it out by themselves.

Stage 2: Standard setting judgement

Before the standard-setting meeting, each panelist was given a set of student oral performances from Grade 4 and Grade 6 respectively. Since the sample sizes were relatively small, with altogether 107 samples from Grade 4 and 103 samples from Grade 6. Accordingly, the samples from each grade level were divided into 12 sets, with 8 to 9 samples in each set, and given to each panelist for judgement. Each set included a random selection of one sample from each decile in the distributions of the total scores, chosen to ensure that the oral performance represented in the sample would also be representative of the population of examinees. The samples in each set were not ordered by the test scores and the panelists were unaware of examinees' actual test scores while making their judgements. During the standard-setting meeting, the panelists were asked to listen to the students' oral samples and classify them into 2 categories, qualified or unqualified. Specifically, they were asked to judge whether the Grade 4 students' oral performances fell below Level 1 or were at Level 1 (or above) and whether the Grade 6 students' oral performances fell below Level 2 or were at Level 2 (or above). Their

judgements were then collected, and distributions of total test scores for each of the 2 categories were formed. The 2 distributions were plotted and analysed to arrive at possible cut-off scores (Cizek & Bunch, 2007).

3.5 Data collection procedures

Mixed methods were used and both qualitative and quantitative data were collected for the current study. Qualitative data was collected via classroom observation, student focus groups and teacher interviews. Quantitative data consisted of student and teacher questionnaires and individual student assessments. These different data sources were brought together for triangulation, which aims to strengthen reliability and internal validity (Creswell & Clark, 2007; Merriam, 1998). Data collection for the main study took place from March 2017 to July 2017. The data collection procedures of each instrument are described below.

3.5.1 Data collection procedure for questionnaires

The teacher questionnaire was distributed at the end of the academic year in July 2017. With the help of the district-level ELT coordinator, a link to the online questionnaire was distributed to teachers in all the state primary schools in one local district in Luoyang, China. A brief statement on the study's purpose and ethical issues was included at the beginning of the questionnaire. The EFL teachers were encouraged to participate, however, participation was voluntary and anonymous. The online teacher questionnaire closed in early August 2017.

The student questionnaire was distributed to 4 participating classes in School A and School B at the end of the academic year in late June 2017. The paper version of student questionnaire was administered by the researcher during an English class. The researcher briefly talked about the study aims and gave instructions on how to fill out each section in the questionnaire. It took approximately 20 minutes for students to complete the questionnaire.

3.5.2 Data collection procedure for classroom observation

During classroom observations, the researcher took a non-participating role and sat at the back of the classroom. The teacher was videotaped for the duration of the session. Fieldnotes were also taken simultaneously during the classroom observation in order to cover those aspects of the classroom not captured by the video recordings. Although observation is reviewed as a research process that captures live or authentic data in an unobtrusive manner in a natural setting, participants may behave in a different way when they are being observed (Cohen, Manion, & Morrison, 2007, p. 396). In order to minimize reactivity, the whole teaching unit was observed, in which three lessons were observed in

total for 1 class within a span of 2 weeks wherein each lesson lasted for 40 minutes (Cohen, Manion, & Morrison, 2007, p. 410). In addition, such prolonged observation enabled the researcher to observe regular teaching rather than demo teaching[13]. Classroom observations in School A and School B were scheduled in early April 2017 and early May 2017 respectively. A total of 12 lessons by 4 teachers were observed.

3.5.3 Data collection procedure for interviews

Teacher interviews were conducted shortly after the classroom observations, in mid-April (School A) and mid-May (School B). It enabled the researcher to develop an in-depth understanding of the participants' perspective. Four observed teachers were interviewed individually in their own offices. The purpose of study was stated, and consent forms were signed before the interviews began. Interviews were audio recorded with participating teachers' permission and were conducted in Chinese. Each teacher interview lasted 55 to 60 minutes, yielding about 4 hours of interview data.

The student focus group interviews were also scheduled after the classroom observation phase. For each class, 2 groups of students were interviewed separately in a school meeting room: six high-achievers and 6 low-achievers. They were classified based on their EFL teacher's recommendations and classifications were confirmed by the researcher's classroom observation. However, the participating students were unaware of their classification. Approval to participate was obtained from teachers and parents prior to the interview. During the focus group interview, students were informed of the study purpose and they were encouraged to express their feelings about EFL learning at school freely. In addition, they were also informed that their names were anonymised for reasons of confidentiality. The interviews were conducted and audio-recorded in the children's native language, Chinese, for natural communication. Each interview session lasted about 30 minutes. There were altogether 8 groups, comprising 48 students from the 4 observed classes in School A and School B, generating a total of about 4 hours of interview data.

3.5.4 Data collection procedure for individual assessments

The individual assessments were scheduled at the end of the academic year, after the classroom observation, student questionnaire and focus group interviews. This timeline (1) allowed the students to become familiar with the researcher and (2) ensured that the students have finished the year's lessons.

[13] Demo lessons refer to practice teaching conducted by the teacher in front of their colleagues or teachers from other schools. This is a common teacher development activity in mainland China. Teachers normally prepared one lesson for peer observation and discussion. However, this kind of lesson was not representative of normal teaching practice.

Individual assessments were administered during class time in a classroom to ensure the environment was familiar to the students. Participating students were called one by one for the assessment. During the assessment, Grade 4 students were arranged into small groups and tested in turns to avoid the tendency towards fatigue noted during the pilot study. Grade 6 students were tested in one go. After the assessments, each student was awarded a certificate of completion. The BAS-II Verbal Comprehension assessment took around 10 minutes for each student. The frog storytelling task allotted five minutes for students to get familiar with the whole story, after which students normally took between 5 and 10 minutes to finish the storytelling task. The storytelling task was audio-recorded for later transcription and coding.

3.6 Data analysis

After the data collection, this section describes the quantitative data analysis of teacher and student questionnaire and student individual assessments. It then presented the qualitative data analysis of interview and classroom observation.

3.6.1 Data analysis of questionnaire

The teacher and student questionnaire datasets were screened before any statistical analysis was conducted. A careful examination of data revealed that there was very little data missing. In teacher questionnaire data set, out of 104 responses in total, there was one case were excluded for choosing the same answer (i.e., strongly disagree) in Section B and C. In student questionnaire data set, of all the 222 Grade 4 and Grade 6 students, only one Grade 4 student (2 missing items in section B) and two Grade 6 students (1 student with 4 missing items in section C and another student with 3 missing items in section B) had missing data in their response. Therefore, in order to make full use of the available data set, all the missing items were replaced with the average means of the corresponding variables, which did not distort or change the normal distribution in the later data analysis (Tabachnick & Fidell, 2007).

Both teacher and student questionnaires were analysed at the level of descriptive and inferential statistics, using SPSS version 23.0. At the descriptive level, the means, standard deviations, the percentages, and ranges were calculated, aiming to explore the data collected and to have an overall picture of EFL teaching and learning in the target context.

At the inferential level, principal component analysis with varimax rotation was used to explore the underlying factors that might be associated with teachers' curriculum implementation practices and students' affective attitudes in the target context. The rationale of choosing this type of analysis was as follows: (1) Variables included in the current teacher and student questionnaire were those that were

frequently mentioned in previous studies (Bulter, 2004; Cheung & Wong, 2012; Enever, Lindgren & Ivanov, 2014; Gorsuch, 2000; Rixon, 2013; Wang, 2006). However, these variables were identified in different contexts and might not be directly applicable in the current context. This study, with the aim of being exploratory in nature, employed principal component analysis and tried to regroup and reduce the variables into a smaller set of factors. This type of analysis might help with easily interpreting and understanding the internal structures or constructs in the current context (Field, 2009). (2) Principal component analysis with varimax rotation was also employed by previous exploratory studies (Lasagabaster, 2011; Wang, 2006). Due to the multidimensional character of teachers' perceptions and practices of curriculum implementation, as well as students' affective attitudes towards EFL learning, the use of varimax rotation in the current study attempted to have each variable loaded onto as few factors as possible. However, it should be noted that the limitation of employing varimax rotation to limit cross-loadings of questions across factors was to treat each underlying factor as uncorrelated, and these factors might be correlated to some extent, therefore this assumption should be examined in further research (Allen, 2017; Preacher & MacCallum, 2003).

In applying the principal component analysis, first, the Kaiser-Meyer-Olkin (KMO) measure was used for measuring the sampling adequacy for further factorial analysis. The KMO value was calculated for all and individual variables and the acceptable value was set at greater than .50 (Field, 2009). Second, following Kaiser's criterion, factors with eigenvalues greater than one were retained and interpreted. Third, for simple interpretation, within each factor scale, all the factor loadings with values less than the cut-off value of .30 were suppressed (Field, 2009). Given the factor solution provided by the principal component analysis, the content of variables that were highly loaded onto the same factor was examined for the common themes and each factor scale was then labelled. In addition, the reliability coefficients were measured by Cronbach's alpha separately for each of the factor scales to assess the level of internal consistency. The mean and standard deviation of each factor scale was calculated to get the overall intensity of the participants' responses.

3.6.2 Data analysis of interview

Interview data was analysed via thematic analysis using NVivo 12. All the raw data was first transcribed and examined based on the research questions. Since the interview was conducted in Chinese, the interviewee's native language, the raw data was initially transcribed verbatim into Chinese and member check was used to ensure that interviewees' views were expressed accurately and fairly by inviting the teacher participants to read and comment on the transcripts (Erlandson, Harris, & Skipper, 1993). The Chinese transcriptions were then entered into NVivo 12 for assisting interpretation and coding. Through an inductive analysis approach, at the initial stage, descriptive codes were generated based on the emerging patterns and categories from the raw data instead of a predetermined model or an existing

framework (Patton, 2014). At the next stage of abstraction, the identified categories were further linked and integrated to each other and salient themes emerged through systematic and recurrent comparisons (Miles & Huberman, 1994). Within each theme, sub-themes were generated from the commonalities and differences among four EFL teachers' interview data by constant comparison and contrasting. Some illuminating extracts were selected to be translated into English and were used as the exemplified quotes in the result chapters. A Chinese PhD student in Education with advanced English proficiency was invited to revise the translated extracts based on the Chinese versions for the accuracy of the translation.

3.6.3 Data analysis of classroom observation

The procedures for carrying out the data analysis on the classroom observation data is described below.

First, following the key-three-step criteria, the classroom-based oral and aural assessment activities were identified.

Second, time spent on each activity was calculated against the total amount of class time.

Third, the adapted COLT scheme was coded by placing check marks into the appropriate box under the seven main features.

The whole procedures were conducted by watching the classroom video recordings repeatedly after the observation instead of real time. Because some categories need time to make the judgement. During the course of single activity, one or several categories might be checked off. It involved with three cases: exclusive focus, primary focus and combination focus (Spada & Fröhlich, 1995). One check mark means an exclusive focus on one category only during the activity. One circled check mark indicated the primary focus when more than one category was involved but most of the time was spent on a certain category. The combination focus happened when the same amount of time and emphasis was spent on two or more categories and a check mark was placed under a separate category called combination. Then the time spent on each category was grouped and the percentage against the total amount of assessment time was calculated.

3.6.4 Data analysis of assessments

The students' individual assessment data were analysed at the level of descriptive and inferential statistics, using SPSS version 23.0. At the descriptive level, descriptive statistics were used to understand the nature of the data collected. The means, standard deviation, range, skewness and kurtosis were calculated (Field, 2009).

At the level of inferential statistics, in accordance with the data collected at a continuous level, the independent sample t-test was used to compare the mean difference between Grade 4 and Grade 6 students' test scores to see if change or growth took place over two years of EFL learning at school. Effect size was calculated using Cohen's *d*, and conventional benchmarks were set for the *d* values (i.e., small effect size: d = 0.2, medium effect size: d = 0.5 and large effect size: d = 0.8).

3.7 Ethical considerations

It was important to take ethical issues into consideration before the research was carried out. The University of Oxford has strict ethical guidelines pertaining to research with young people and their teachers in schools, which is consistent with the ethical codes set forth by the British Educational Research Association (2011). A proposal for the current research project was submitted the University of Oxford Central University Research Ethics Committee (CUREC). The CUREC approval was obtained before both the pilot and main data collection phase.

Before the commencement of this study, school principals in China were approached via email with comprehensive information about this study (Appendix 10). Both teachers and students involved in this study were fully informed of the study's purpose; the procedures involved in the research; the time required to complete a number of certain tasks (e.g., individual tests, questionnaire, interview), the possible risks or benefits; their rights to drawn from the study at any time point; the way of storing and managing with the data and the way of disseminating the findings (Appendix 11; Appendix 12). The involving participants were then asked to sign a consent form (Appendix 13). As the students in the current study were under the age of 18, they were regarded as 'people whose ability to give free and informed consent is in question' (BERA, 2011). Accordingly, informed consent for student participants was obtained from parents or guardians for their children's participation (Appendix 14). Parents and guardians were specifically informed that they and their children have the right to withdraw from the research for any or no reason, and at any point in time, with no consequence. Teachers and students' parents or guardians were also informed that all data collected were kept strictly confidential. Children were allowed to conduct tasks at their own pace and opportunities were provided for them to share their views or raise any questions, reducing the likelihood of distress or discomfort. Burdens on participants were minimised by ensuring that each task was kept short and child friendly. Each participant was given an identification number in the record forms.

In terms of the confidential issues, pseudonyms and codes were used for both teacher and student participants during data collection, data analysis and reporting results periods. The name of participating teachers was assigned with random capital letters, for example, Ms P and Ms W; students were allocated with alphanumerical codes, for example, Student 1 was written as S1. All the data were kept in a secure

file to which only the researcher had access over the course of this research. This could ensure that the participants could not be identified.

3.8 Summary of chapter three

This chapter discusses methodological issues in the current study. First, it describes the mixed-methods approach employed in the study and presents overview of the research design. Then the sampling frame and the criteria for the sample selection are discussed and followed by demographic information about the participants. The rationale for each instrument used for the current study is discussed, and adaptations to the instruments described. The data collection procedures and data analysis are described, followed by issues related to ethical considerations.

Chapter 4 Data Analyses and results: classroom context

4.1 Introduction

This chapter reports on research findings in the classroom context, which correspond to the first main research question:

RQ1. How do Grade 4 and Grade 6 EFL teachers in state primary schools conduct their assessments in listening and speaking in the classroom contexts in relation to the NECS targets?

> RQ1.1 Which classroom-based assessment activities do EFL teachers use for students in Grade 4 and Grade 6?
>
> RQ1.2 How are these assessment activities implemented in the classroom?

Classroom data is reported before the discussion of teachers' attitudes toward the NECS targets and classroom-based assessment, as well as before students' learning outcomes, because knowledge of classroom activities and practices, as the first layer of investigation, facilitates understanding of the target context. Then, with this chapter having discussed what was observed in the classroom, the next chapter reports factors affecting the way teachers implement the curriculum as well as their attitudes towards NECS targets and classroom-based assessment.

4.2 Teacher's commonly used classroom-based assessment activities in assessing listening and speaking skills

This section addresses the research question of RQ1.1, *Which classroom-based assessment activities do EFL teachers use for students in Grade 4 and Grade 6?*, to find out which classroom-based oral and aural assessment activities were used by primary EFL teachers and how frequently these assessment activities were carried out in the classroom. Since this section aims to set the general scene prior to presenting more in-depth, detailed analysis, findings from a teacher questionnaire of a total of 103 EFL teachers from 28 state schools of the local district are reported first. Next, results of classroom observations of four EFL teachers' daily classroom teachings are reported.

4.2.1 Results of the teacher questionnaire

This section presents the findings for section C1 in the teacher questionnaire (see Appendix 2). Teacher participants were asked to indicate the frequency of using 18 classroom-based assessment activities in their daily classroom teachings on a 5-point Likert scale (i.e., 5 = Always: in almost every lesson; 4 = Frequently: once a week; 3 = Some time: 2 to 3 times a term; 2 = Rarely: once a term and 1 = Never). These activities assess young EFL learners' skills ranging from unidirectional listening and speaking to

integrated listening and speaking skills. The mean frequency of each classroom-based assessment activity was calculated, and the rankings are depicted in Table 4.1.

Table 4. 1 The frequency-ranked list of classroom-based assessment activities used by teachers in assessing listening and speaking skills

Rank	Type of assessment activities	Mean (SD)	Rounded Mean	Targeted skills
1	Read aloud	4.83 (0.42)	5	Pronunciation and intonation
2	Listen and imitate	4.79 (0.53)	5	
3	Q&A led by the teacher	4.77 (0.55)	5	Understanding and responding to questions
4	Listen and answer	4.76 (0.47)	5	
5	Listen and act	4.71 (0.51)	5	
6	Recite text/dialogues	4.64 (0.57)	5	Memorisation at discourse level
7	Textbook role-play	4.42 (0.62)	4	Communicative skills
8	Textbook retelling	4.39 (0.72)	4	Narrative skills
9	Q&A led by the group/pairs	4.38 (0.94)	4	Understanding and responding to questions
10	Make up conversation	4.37 (1.04)	4	Communicative skills
11	Dictation	4.36 (0.74)	4	Memorisation at vocabulary level
12	Make up sentences	4.35 (0.64)	4	Producing/completing fix-patterned single sentences
13	Oral gap filling	4.32 (0.77)	4	
14	Oral translation	4.22 (0.94)	4	Translation skills
15	Listening comprehension (in the form of paper-and-pencil exercises)	4.06 (0.94)	4	Listening comprehension skills
16	Picture description	3.29 (1.02)	3	Narrative skills
17	Storytelling	2.82 (0.92)	3	
18	English drama	2.34 (0.62)	2	

As identified in the table, the most frequently used classroom-based assessment activities were the following: read aloud, listen and imitate, Q&A led by teacher, listen and answer, listen and act and recite text/dialogue. The rounded mean of these activities was 5, which indicated that these activities were used in almost every lesson by the surveyed teachers. Among these 6 activities, the top two were form-focused activities directed toward assessing students' pronunciation and intonation, in correspondence with the first NECS target, which focuses on speaking skills at two levels: students were expected to imitate from the recordings (Level 1) and to articulate English clearly with the appropriate intonation (Level 2). The following 3 activities (i.e., Q&A led by teacher, listen and answer and listen and act) involved both listening and speaking skills, meaning that students were required to give either verbal or physical responses based on what they heard. These activities were matched with NECS targets focused on assessing students' understanding of simple questions, instructions or short stories. The last activity that was reported as being used in almost every lesson was recite text/dialogue. This activity mainly focused on checking students' memorisation at the discourse level.

In the second part of the ranked list, there were 9 activities with the rounded mean of 4, indicating that they were used around once a week by the participating teachers. Among these activities, the first two

(i.e., textbook role-play and textbook retelling) used cued language from the textbook to assess students 'communicative and narrative skills; these skills partially correspond to the NECS targets emphasising students' expected ability to produce expressive and narrative phrases and sentences. The next two most frequent activities with a rounded mean of 4 (i.e., Q&A led by group/pairs and make up conversations) were focused on assessing students' understanding of and responses to their peers' questions. The remaining four activities (i.e., dictation, make up sentences, oral gap filling and oral translation) were mainly form-based and assessed students' use of vocabulary and sentences in an isolated context. Listening comprehension was the last activity that was reported as having been frequently used by the teachers, although the assessment of students' comprehensive listening skills was completed in the form of pencil-and-paper exercises.

The least frequently reported assessment activities were picture description, storytelling and English drama. Picture description and storytelling had a rounded mean of 3, indicating that they were reported to be used 2 to 3 times per term. English drama was reported to be rarely used, with the rounded mean of 2 (i.e., once per term). These activities assessed students' ability to use their own language (i.e., without cued language from the textbook) to express their ideas and to narrate at a discourse level.

4.2.2 Results of classroom observations

This section presents findings from classroom observations of four EFL teachers' lessons in Grade 4 and Grade 6. Each teacher was observed teaching for a whole unit, which consisted of three lessons in total. The section first provides an overview of which classroom-based assessment activities were conducted in the teachers' actual classrooms from two grade levels, and then relevant teacher interview data is reported to triangulate the findings obtained from classroom observation.

4.2.2.1 Classroom-based assessment activities used in Grade 4

In this section, the sequence of discussion is as follows: first, the classroom-based assessment activities used by each of the two observed Grade 4 EFL teachers are summarised; next, some details regarding how these activities were organised (i.e., teacher-centred or learner-centred) and what the intended content focus of these activities were (form-focused, meaning-focused or a combination of form and meaning focused) are discussed; then, finally, which skills were targeted (listening only, speaking only or integrated skills) and which kind of materials were used (textbook, audio or visual) are explored. The commonalities and differences between the two teachers are synthesized, aiming to provide a relatively comprehensive picture of the use of assessment activities in the sampled Grade 4 classrooms.

Types of classroom-based assessment activities

Following the key three-step criteria (see Section 3.4.3.1 for a detailed description of how assessment activities in classrooms were identified), two EFL teachers' classroom-based assessment activities were identified from a detailed analysis of classroom video transcripts. Table 4.2 summarises the types of assessment activities used in the two teachers' lessons. The percentage represents the proportion of class time spent on each type of assessment activity. The results revealed that, in general, of the three observed lessons, with a total of 120 minutes of instruction for each teacher, around half of the class time was spent engaged in assessment activities, with 58.8% in Ms P's class and 43.3% in Ms W's class. In addition, both teachers conducted a variety of assessment activities in their classes, with 10 assessment activities identified in Ms P's class and 8 in Ms W's class.

There were 7 common activities found in both teachers' classes: Q&A led by teacher; listen and answer; read aloud; make up conversations; textbook retelling; textbook role-play and Q&A led by group. Among these, the top 3 assessment activities with the most time devoted to them by both teachers overlapped. These activities were Q&A led by teacher, listen and answer and read aloud, indicating that these activities were the most frequently used to assess students' listening and speaking skills in the sampled Grade 4 classrooms. In terms of their correspondence to NECS targets, both Q&A led by teacher and listen and answer activities were used to assess students' understanding of and response to their teacher's questions; these assessments corresponded to NECS targets in communicative skills at Level 1. Additionally, read aloud was used to assess students' pronunciation and intonation. In contrast, the activities with the least time devoted to them were textbook retelling; textbook role-play and Q&A led by group. These activities required students to have a longer output comprised of whole sentences, and sometimes to do so in a narrative manner, in correspondence with NECS targets in narrative skills at Level 1. In sum, targets related to communicative skills were most frequently assessed by teachers in the classroom. However, targets related to narrative skills were less frequently assessed.

Table 4. 2 The frequency-ranked list of two Grade 4 teachers' classroom-based assessment activities

	Ms P, School A			**Ms W, School B**		
Rank	**Activity**	**Time**	**%**	**Activity**	**Time**	**%**
1	Q&A led by teacher	19 min	15.8%	Q&A led by teacher	15 min	12.5%
2	Read aloud	14.5 min	12.1%	Listen and answer	7.5 min	6.3%
3	Listen and answer	10 min	8.3%	Read aloud	7 min	5.8%
4	Q&A led by group	9.5 min	7.9%	Make up conversation	6.5 min	5.4%
5	Textbook retelling	5 min	4.2%	Textbook retelling	6 min	5%
6	Make up conversation	4 min	3.3%	Textbook roleplay	4 min	3.3%
7	Make up sentences	3 min	2.5%	Q&A led by group	3.5 min	2.9%
8	Listen and act	2 min	1.7%	Oral gap filling	2.5 min	2.1%
9	Textbook roleplay	2 min	1.7%			
10	Listen and imitate	1.5 min	1.3%			
Total		70.5 min	58.8%		52 min	43.3%

Note: % indicates the percentage of total class time, which consisted of 3 40-minute lessons, for 120 minutes total, for each teacher

Main features of classroom-based assessment activities

Four main features of the assessment activities utilized in the two teachers' lessons are documented in the adapted COLT scheme and are presented in Table 4.3. The percentage represents the proportion of total assessment activity time spent on the individual categories under each of the four features. It aimed to explore how these identified assessment activities in listening and speaking skills were related to NECS targets regarding the activities' participant organisation, intended content focus, student modality and materials used.

Table 4. 3 Time and percentage of time spent on different categories by two Grade 4 teachers

	Ms P, School A		**Ms W, School B**	
Features	**Time**	**%**	**Time**	**%**
Participant Organisation				
Whole class	52 min	73.8%	36 min	69.2%
Group	18.5 min	26.2%	16 min	30.1%
Individual	0 min	0 %	0 min	0 %
Intended content focus				
Form	16 min	22.7%	9.5 min	18.3%
Meaning	43.5 min	61.7%	31.5 min	60.1%
Combination	11 min	15.6%	11 min	21.2%
Student modality				
Listening	2 min	2.8%	0 min	0%
Speaking	17.5 min	24.8%	9.5 min	18.3%
Integrated	51 min	72.3%	42.5 min	81.7%
Type of material				
Written text minimal	14.5 min	20.6%	7 min	13.5%
Written text extended	42.5 min	60.3%	31.5 min	60.6%
Audio	13.5 min	19.2%	7.5 min	14.4%
Visual	5 min	7.1%	6 min	11.5%

Note: % indicates the percentage of total classroom-based assessment time

First, *participant organisation* was used to describe whether the classroom-based assessment activities were teacher-centred or student-centred. As NECS targets in speaking and listening skills emphasised developing students' oral language ability, group activities were considered to be essential for developing students' communicative competence; in contrast, teacher-centred activities were considered to be potentially restrictive of students' use of oral language (Spada & Fröhlich, 1995). In Table 4.3, both Grade 4 teachers displayed similar patterns of participant organisation. Teacher-centred classroom-based assessment activities were found to be dominant in both teachers' lessons. Ms P and Ms W spent 73.8% and 69.2% of their total respective assessment time on teacher-centred activities. In contrast, relatively little time was spent on group assessment activities, which comprised of only 26.2% (Ms P) and 30.1% (Ms W), respectively, of the total assessment time. No individual assessment activities were observed.

Second, the *intended content focus* aimed to investigate whether the primary focus of the classroom-based assessment activities was on form and/or meaning. Focusing on form indicated that the teachers assessed students' linguistic forms, such as vocabulary, grammar or pronunciation, instead of meaning. To be clear, NECS targets prioritized the assessment of the meanings conveyed over the form of the

language. Table 4.3 shows that the majority of both teachers' assessment time was devoted to meaning-focused assessment activities, 61.7% in Ms P's total assessment time and 60.1% in Ms W's. The time allocated to form-focused assessment activities was much less (22.7% and 18.3%). In addition, there was also a small portion of time devoted to activities combining assessment of both form and meaning (15.6% and 21.2%).

The third category, *student modality*, identified which kinds of skills were involved in the classroom-based assessment activities. It was found that both Grade 4 teachers devoted most of their assessment time to activities that integrated listening and speaking skills (72.3% and 81.7%, respectively). This finding also corresponded with the finding in the teacher questionnaire (Section 4.2.1) that activities like Q&A led by teacher and listen and answer were reported to be the most frequently used assessment activities in primary classrooms. In contrast, single skill (listening or speaking) involved activities were relatively less frequent during assessment.

The last feature, *type of material*, was intended to describe which kinds of materials was used during the classroom-based assessment activities. The results showed that written text materials (80.9% and 74.1%) were used predominantly by both teachers. In addition, both teachers used extended text (60.3% and 60.6%) such as dialogues and connected sentences more frequently than they did minimal text (20.6% and 13.5%) such as the use of isolated sentences or of a vocabulary list. Audio and visual materials were also found to be used by the two teachers, composing 18.5% in Ms P's total assessment time and 13.5% in Ms W's. Based on classroom observations, the written text and audio materials were all textbook-oriented.

In sum, both of the Grade 4 teachers displayed a similar pattern regarding the features of their classroom-based assessment activities: Teacher-centred, meaning-focused, extended written text materials with textbook-oriented activities integrating both listening and speaking skills were mainly used by both instructors. These activities seemed to be congruent to NECS targets in listening and speaking skills focused on demonstrating students' authentic use of oral language in a meaningful context.

Relevant data from teacher interviews

In this section, findings from the interviews with the two Grade 4 teachers are presented both to supplement and to triangulate with the previous questionnaire and classroom observation data. During the interviews, teachers were asked about the most frequently used types of classroom-based assessment activities during ordinary lessons. Commonalities and differences within the interview data are summarised with multiple quotations.

Five types of commonly used classroom-based assessment activities were mentioned by both teachers for the assessment of students' listening and speaking skills. These activities were Q&A led by teacher, textbook role-play, textbook exercises, dictation and recite text/dialogue from the textbook. The first two activities targeted integrated listening and speaking skills while also combining form-focused and meaning-focused assessment. The Q&A led by teacher activity was further divided into two forms: one was restricted Q&A, in which the teacher posed display questions and expected ready-made answers from students; the other was personalised Q&A, in which the teacher asked questions regarding students' real lives or true feelings. As the following quotes illustrate,

> In my classroom, the assessment activities I often used were those which engaged my students in the integrated skills areas...So, I always use [restricted] Q&A to assess my students on whether they could understand the target sentence patterns and respond to the given questions using target words. This activity always comes first before the free Q&A between teacher or peers, which I use to assess whether they could apply the target sentence in their real life. (Ms P, Grade 4, School A)

> I prefer using textbook role-play to assess my students' oral skills during the class. It normally involves 3 to 4 students, and they were asked to perform the role-play using the text or dialogue sentences from the textbook. Everyone got a chance to speak, unlike the choral answers. Both the whole class and I could evaluate each student's performance later on. (Ms W, Grade 4, School B)

The other three activities also mentioned by both teachers were either paper-and-pencil based or form-focused and memorisation-based. Although these activities were seldom observed by the researcher during classroom observation, they were treated as regular assessment activities by the two teachers. As Ms W further explained,

> I often use textbook exercises as the class quiz to assess students' listening comprehension skills when I finish teaching one or two units. These exercises were usually in the multiple-choice format. I also use dictation, probably once every two weeks, to check students' memorisation of heard vocabularies. All these activities could be seen as the preparation and foundation for the exams. (Ms W, Grade 4, School B)

In sum, the teacher interview data revealed that Grade 4 students' listening and speaking skills were frequently assessed through both contextualised and context-reduced assessment activities. The contextualised assessment activities (e.g., Q&A and textbook role-play) involved either teacher and student interaction or peer interaction with meaning expressed. However, context-reduced activities, also known as "stand alone activities" were less interactive. They involved evaluating students' memorisation at the lexical and discourse levels (e.g., recite text/dialogue and dictation) or were administered in 'exam-like' conditions (e.g., textbook exercises). Neither type of activity had been observed in class but were recognised by both teachers (Cummins, 1984; Rea-Dickins, 2001). This finding was also in line with results from the previous teacher questionnaire, which discovered that form-focused assessment activities were claimed to be frequently adopted by local teachers.

4.2.2.2 Classroom-based assessment activities used in Grade 6

This section follows the same presentation structure as in Section 4.2.2.1; however, in this section, the focus is on Grade 6. The types of assessment activities used in two Grade 6 EFL teachers' classrooms are presented, followed by an analysis of the activities' associated main features. Supplemental data from interviews with both teachers is provided to be triangulated with data from classroom observations.

Types of classroom-based assessment activities

Following the key three-step criteria, two Grade 6 teachers' classroom-based assessment activities were identified from an analysis of classroom video transcripts. Table 4.4 summarises the types of assessment activities in the two teachers' lessons. The percentage represents the proportion of class time spent in each type of assessment activity. In general, of the three observed lessons, with a total of 120 minutes for each teacher, more than half of the class time (57.9% in Ms L's and 66.3% in Ms B's class) was spent engaged with assessment activities.

Both teachers employed a variety of assessment activities in their classes. Eight assessment activities were identified in Ms L's class and 11 in Ms B's. Seven assessment activities found in common in both teachers' classes. These were Q&A led by teacher; textbook retelling; Q&A led by group; oral gap filling; listen and answer; read aloud and textbook role-play. Among these, Q&A led by teacher ranked highest in Ms L's classroom listen and answer ranked highest in Ms B's; these activities constituted 16.3% and 15.4% of each teacher's total class time, respectively. The two activities integrated listening and speaking skills and were used to assess students' understanding of and response to the teacher' questions in class. The second and third most frequent types of assessment activities in both teachers' classrooms overlapped. They were textbook retelling and Q&A led by group, both targeting students' narrative and communicative skills. In contrast, read aloud and textbook role-play were less frequently used activities in both teachers' classes. These types of activities were used by teachers to check the students' pronunciation and their rote learning of textbook dialogue.

In addition, some form-focused, memorisation-based activities (e.g., make up sentences, dictation and recite text/dialogue) were found in Ms B's class. Though the focus of these activities was not congruent with NECS targets emphasising students' oral language use ability, the findings were also confirmed by the results of the teacher questionnaire, which found that these same activities were reported as frequently used by local teachers.

Table 4. 4 The frequency-ranked list of two Grade 6 teachers' classroom-based assessment activities

	Ms L, School A			Ms B, School B		
Rank	**Activity**	**Time**	**%**	**Activity**	**Time**	**%**
1	Q&A led by teacher	19.5 min	16.3%	Listen and answer	18.5 min	15.4%
2	Textbook retelling	13.5 min	11.3%	Q&A led by group	14 min	11.7%
3	Q&A led by group	9 min	7.5%	Textbook retelling	13 min	10.8%
4	Oral gap filling	8 min	6.7%	Q&A led by teacher	9.5 min	7.9%
5	Listen and answer	7 min	5.8%	Listen and imitate	5 min	4.2%
6	Read aloud	5 min	4.2%	Textbook role-play	4.5 min	3.8%
7	Listen and act	4 min	3.3%	Oral gap filling	4 min	3.3%
8	Textbook roleplay	3.5 min	2.9%	Make up sentences	3.5 min	2.9%
9				Recite text/dialogue	3 min	2.5%
10				Dictation	2.5 min	2.1%
11				Read aloud	2 min	1.7%
Total		69.5 min	57.9%		79.5 min	66.3%

Note: % indicates the percentage of total class time, which consisted of 3 40-minute lessons, for 120 minutes total, for each teacher

Main features of classroom-based assessment activities

Four main features of the assessment activities used in the two Grade 6 teachers' lessons were captured in the adapted COLT scheme and are presented in Table 4.5. The percentage represents the proportion of total assessment activity time spent on the individual categories under each of the four features. The aim of the table is to explore how the identified assessment activities in listening and speaking skills relate to the NECS targets regarding the features of activities' participant organisation, intended content focus, student modality and materials used.

Table 4. 5 Time and percentage of time spent on different categories by two Grade 6 teachers

	Ms L, School A		Ms B, School B	
Features	**Time**	**%**	**Time**	**%**
Participant Organisation				
Whole class	57 min	82%	51.5 min	64.8%
Group	12.5 min	18%	28 min	35.2%
Individual	0 min	0 %	0 min	0 %
Intended content focus				
Form	7 min	10.1%	4.5 min	5.7%
Meaning	25.5 min	36.7%	13.5 min	17%
Combination	37 min	53.2%	61.5 min	77.4%
Student modality				
Listening	4 min	5.8%	2.5 min	3.1%
Speaking	26.5 min	38.1%	25.5 min	32.1%
Integrated	39 min	56.1%	51.5 min	64.8%
Type of material				
Written text minimal	8 min	20.6%	6 min	7.5%
Written text extended	48.5 min	60.3%	37 min	46.5%
Audio	7 min	19.2%	23.5 min	29.6%
Visual	4 min	7.1%	13 min	16.4%

Note: % indicates the percentage of total classroom-based assessment time

First, the results under the category of *participant organisation* indicated that the activities that were teacher-centred made up the largest proportion of assessment time in both teachers' lessons, which constituted 82% (Ms L) and 64.8% (Ms B) of their total assessment time, respectively. Consequently, time devoted to group-centred assessment activities was relatively little (18% and 35.2%). No individual assessment activities were identified.

Second, the *intended content focus* in both teachers' assessment activities was mainly form-meaning in combination, accounting for more than half of the total assessment time (53.2% and 77.4%). Meaning-focused assessment activities constituted 36.7% in Ms L's classroom and 17% in Ms B's classroom of total assessment time. By contrast, the time allocated to the direct assessment of language forms was less (10.1% and 5.7%). In comparison with the intended content focus of Grade 4 assessment activities, which mainly focused on assessing the meaning expressed, Grade 6 activities targeted both language structure and intended meaning.

Third, with regard to *student modality*, both Grade 6 teachers' lessons focused on listening and speaking integrated skills (56.1% and 64.8%), much as both Grade 4 teachers' lessons did. The percentage for

assessing speaking only skills activities was relatively lower, 38.1% and 32.1% in Ms L's and Ms B's lessons, respectively. Very few activities assessing listening skills only were found in either teacher's class (5.8% and 3.1%).

Turning to the last feature, *type of material*, it was found that both Grade 6 teachers primarily used written text materials to assess students' listening and speaking skills, though Ms L (80.9%) used such materials far more frequently than Ms B (54%) did. However, both teachers more frequently used extended text materials (60.3% and 46.5%) than minimal forms (20.6% and 7.5%). Audio and visual materials were also found to be used by both teachers. Specifically, Ms B spent nearly half of the assessment time (46%) using these authentic materials while Ms L allocated relatively less time to them (26.3%). Based on classroom observations and in keeping with the findings for Grade 4, both written text and audio and visual materials were textbook-oriented.

To sum up, the features of both Grade 6 teachers' assessment activities were mostly similar: teachers acted as the main organiser of the classroom-based assessment activities; form-meaning combined assessment activities were frequently used to assess students' integrated oral language use ability; and the sources of assessment activities were, predominantly, textbook-oriented materials in both written and audio forms.

Relevant data from teacher interviews

In this section, findings from two Grade 6 teachers' interviews are presented both to supplement and to triangulate with the questionnaire and classroom observation data. During the interviews, teachers were asked about the most frequently used types of classroom-based assessment activities in their typical lessons. Commonalities and differences within the interview data were summarised with multiple quotations.

Four types of commonly used classroom-based activities for assessing students' listening and speaking skills were mentioned by both Grade 6 teachers. There were Q&A led by teacher; Q&A led by group, textbook retelling and recite text/dialogue. The first two activities targeted the assessment of language use at two levels: (1) students' mastery of pre-learned sentence patterns and/or vocabulary; and (2) students' ability to use target sentences or vocabulary words in a real-life context. Unlike the Q&A assessment activities in Grade 4, which mainly involved eliciting and assessing students' one-word responses or formulaic expressions, the Q&A assessment activities in Grade 6 tended to focus on information gap or information exchange related tasks (e.g. survey, personalised questions) to get students to communicate or to use target language for an actual purpose. As Ms B further illustrated,

> Since my students were in a higher-grade level, I think the Q&A activities involved with inform-gap tasks were more meaningful and attractive for them. For example, when I want to assess whether or not they could use the sentence pattern like *What do you like?* and *What does he/she like?* in the real-life context, I usually ask my students to do a group survey. They need to ask their group members about what they like and summarize the information in a table and then present [that information] to the whole class. (Ms B, Grade 6, School B)

In addition, both teachers mentioned that Q&A led by the teacher usually comes before the Q&A led by the group. Following this sequence for the two assessment activities not only allowed teachers to interact with and to assess the whole class or individual student, but it also served as a model or as an example for later group assessment work to follow. As Ms L said,

> I think during the class assessment activities, Q&A between teachers and students should always come first before any group activities. I do recognise that peers or group interaction is very important, but they need me to set a model for them; otherwise, the group work might not meet my expectation. (Ms L, Grade 6, School A)

Textbook retelling and recite text/dialogue were also frequently cited by both teachers during their interviews. Both assessment activities required students' language production at a discourse level; however, textbook retelling mainly focused on assessing students' cued narrative skills while recite text/dialogue mainly focused on checking students' memorisation at the discourse level. Although recite text/dialogue was not encouraged in the NECS and was not in line with NECS targets in speaking skills, the assessment activity was still frequently used by both teachers. Ms L offered the following explanation,

> The purpose of reciting the text/dialogue is not just assessing my students' pronunciation, but to encourage students to open their mouth to just say something from the textbook, especially for those low achievers. How can they express themselves or communicate with foreigners if they refuse to open their mouth? Recite is at least a first step for them to speak out. (Ms L, Grade 6, School A)

To sum up, the teacher interview data revealed that oral assessment activities in Grade 6 classrooms mainly involved prompting student responses to questions (which were raised by teachers or by peers) and eliciting students' narrative output by means of textbook retelling and/or recite text/dialogue. Although some of these activities were also found in the Grade 4 class and in the same format (e.g., Q&A), the Grade 6 assessment activities went beyond assessing students' pre-learned vocabulary words or formulaic expressions. Information gap tasks were also set and students were given more opportunities to use their own language or ideas creatively in accordance with the proficiency and cognitive level of Grade 6 students. It seems that the frequently used assessment activities were language-use oriented and covered most of the NECS targets in listening and speaking skills. However, how these assessment activities were implemented still remains unknown. Investigating these implementation methods is the main focus of the next section.

4.3 Teacher classroom-based assessment practices

The previous section mainly looked at which assessment activities were used to elicit students' evidence of learning as well as what features were associated with these activities. This exploration corresponded to the first key step in the classroom-based assessment process (i.e., eliciting students' learning evidence) and was analysed at the activity level.

This section addresses the research question of RQ1.2, *How were these assessment activities implemented in the classroom?* This section is focused on the level of practice. Specifically, this analysis looks at (1) who is actually conducting the assessment (i.e., agent of assessment); (2) what their focus is during the process of assessment and (3) how they make use of their judgement. This process corresponds to the second and third key steps of classroom-based assessment, which are interpreting students' responses and providing feedback. Following the same structure as the previous section's, this section first reports the findings from the teacher questionnaire and then moves on to discuss the results of classroom observations of four EFL teachers' assessment practices.

4.3.1 Results of teacher questionnaire

This section presents the findings from Section C2 in the teacher questionnaire (see Appendix 2). There were two parts in Section C2, agent and focus of assessment practices, are discussed, respectively, in the following sections.

First, teacher participants were asked to indicate the frequency with which they used different evaluators during classroom-based assessment activities; participating teachers were instructed to indicate frequency on a 5-point Likert scale (i.e., 5 = Always: in almost every lesson; 4 = Frequently: once a week; 3 = Some time:: 2 to 3 times a term; 2 = Rarely: once a term and 1 = Never). The evaluators, also called agents of assessment, included teachers, peers and the individual learner (Harlen, 2011). They all might become involved with evaluating students' learning evidence as it is elicited by assessment activities (Black & Wiliam, 2009). In Table 4.6, the mean frequency for each agent of assessment was calculated and ranked. It was found that teacher assessment ranked first, with a rounded mean of 5. This score indicates that the teacher's role as an agent of judgement dominated in almost every lesson. Peer assessment ranked second, with a rounded mean of 4, revealing that this form of assessment was also reported to happen once a week in the local context. In contrast, student self-assessment, with a rounded mean of 3, was relatively seldom used.

Table 4. 6 Agent and focus of assessment practices

Rank		Mean (SD)	Rounded Mean
Agent of assessment practices			
1	Teacher assessment	4.85 (0.42)	5
2	Peer assessment	4.16 (0.53)	4
3	Student self-assessment	3.42 (0.92)	3
Focus of assessment practices			
1	Pronunciation & intonation	4.78 (0.87)	5
2	Lexis	4.23 (0.62)	4
3	Affective attitudes (motivation & confidence)	4.04 (0.43)	4
4	Content/topic	3.77 (0.61)	4
5	Spelling	3.39 (0.94)	3
6	Functional language	3.27 (0.56)	3
7	Grammar	2.67 (0.89)	3

Second, teacher participants were asked to indicate the importance of a number of different types of content focus (e.g., pronunciation and intonation; lexis and affective attitudes) to the assessments of students' listening and speaking skills. A 5-point Likert scale was used (i.e., 5 = Very important; 4 = Fairly important; 3 = Important; 2 = Slightly important; 1 = Not at all important). It can be noted from Table 4.6 that students' correct use of pronunciation and intonation, with the rounded mean of 5, was regarded by the teacher participants as the most important focus area. Also ranked as fairly important by teacher participants were the following: the use of vocabulary students' motivation and confidence in participating in the activities; and the content or topic of the students' response. These results revealed that local teachers also take into consideration the students' expressed meaning and their affective attitudes during the assessment. In contrast, the content areas regarding the accuracy of linguistic forms like correct spelling, the correct use of functional language and the correct use of oral grammar were ranked as relatively less important by local teachers.

Taken together, the data from the teacher questionnaire indicated that teachers act as the main agent during assessment, although peers do also get involved sometimes. Additionally, pronunciation and intonation were considered to be the most important focuses for oral assessment. The content and vocabulary use of students' response and the students' willingness to participate were also taken into consideration by local teachers. In contrast, the specific linguistic forms were considered as much less important during the process of assessment.

4.3.2 Results of classroom observations

This section examines data documented by classroom observation, it is reporting on the actual implementation of assessment activities in the four observed teachers' classes. The section also provides relevant classroom excerpts as examples, as well as teacher interview quotations as supporting details, to be triangulated with the findings from teacher questionnaire.

4.3.2.1 Classroom-based assessment practices in Grade 4

In this section, the results of classroom-based assessment practices by two Grade 4 EFL teachers are presented in Table 4.7. The percentage represents the proportion of total assessment activity time spent on the individual categories listed under each of the three practical dimensions, aiming to provide a relatively comprehensive picture of the assessment practices in the sampled Grade 4 classrooms. The three dimensions include: 1) who is conducting the assessment (i.e., the agent of assessment); 2) who is being assessed (i.e., the target of assessment) and 3) what the actual focus was for judging students' response (i.e., the actual focus of assessment). The commonalities and differences between the two teachers' responses were synthesized.

Table 4. 7 Time and percentage of time spent on different categories by two Grade 4 teachers

	Ms P, School A		**Ms W, School B**	
Features	**Time**	**%**	**Time**	**%**
Agent of assessment				
Teacher	59.5 min	84.4%	46.5 min	89.4%
Peer	11 min	15.6%	5.5 min	10.6%
Target of assessment				
Whole class	13.5 min	19.1%	3 min	5.8%
Group	8.5 min	12.1%	2.5 min	4.8%
Individual student	48.5 min	68.8 %	46.5 min	89.4 %
Actual focus of assessment (feedback)				
Forms	19 min	26.9%	15 min	28.5%
Meanings	19.5 min	27.7%	9.5 min	18.3%
Combination	32.5 min	46.1%	27.5 min	52.9%

Note: % indicates the percentage of total classroom-based assessment time

Agent of assessment

The results indicated that most of the assessments in class were conducted by the teacher, with 84.4% and 89.4% , respectively, of the two teachers' total assessment time directed in this fashion. In contrast, assessment made by peers was far less frequency evidenced (15.6% in Ms P's and 10.6% in Ms W's

classrooms). According to classroom observations, peer assessment was also teacher-led, originating mostly in group activities in which teachers encouraged students to judge their peers' responses by asking questions such as "Any comments?" or "What do you think of their performance?"; the teachers then provided their own feedback afterwards (See Extract 4 in the following section).

Target of assessment

Turing to the target of assessment, it was found that individual students' responses were mostly assessed, taking up 68.8% and 89.4%, respectively, of Ms P's and Ms W's total assessment time. In contrast, choral answers made by the whole class or by a group of students were less frequently assessed in both teachers' classes (19.1% and 12.1% in Ms P's class; 5.8% and 4.8% in Ms W's).

According to classroom observation, both teachers tended to ask for individual volunteers or to nominate students to participate in assessment activities that checked whether the students had achieved the learning targets. As the teachers stated,

> I usually first give the opportunities to the active students who always raise their hands and then nominate some low-achieving students. This would give me the overall idea of whether they could achieve the learning goals. (Ms P, Grade 4, School A)

> I normally ask a row of students to give their response one by one. This way might cover students with mixed ability and would give me a general idea of the extent to which they have mastered the content or not. (Ms W, Grade 4, School B)

Actual focus of assessment: feedback

The actual focus of assessment indicated whether the outcome of assessment activities was evaluated according to the use of linguistic forms, the meaning expressed, or a combination of the two. This finding was reflected by the content of the feedback provided by teachers or peers.

As can be seen from Table 4.7, both teachers spent nearly one-third of their assessment time (26.9% and 28.5%, respectively) focused on assessing the linguistic accuracy of students' responses. According to classroom observations, the teachers' feedback was usually provided in the form of corrections, and the content of the feedback was related to pronunciation, vocabulary and grammar (See Extract 3). Forms-related feedback was also found in most of the peer assessment. Specifically, peer feedback tended to focus more on correcting pronunciation (See Extract 4). The following extracts serve as examples to illustrate this finding.

Extract 3 (Ms P, Grade 4, School A) Teacher-led form-related feedback in a Q&A assessment activity

T:	Now, let's do the Q&A. xxx, please come here. I will ask, and you answer me, OK?
	[Teacher asked one student to come to the front of class]
T:	Ok, listen carefully. Do you like summer?
S:	No, I do.
T:	No, I... DON'T...
S:	OK, no, I don't.
T:	Yes, very good. Now please ask me.
S:	Do you have a spring?
T:	No, you should say, do you LIKE...
S:	Do you like spring?
T:	Yes, I do.

Extract 4 (Ms W, Grade 4, School B) Teacher-led peer form-related feedback in a make-up conversation assessment activity

T:	Now, this group has presented their new made dialogue. Any comments?
S1:	xxx [S2's name], follow me, snowman.
S2:	Snewman (S2's wrong pronunciation).
T:	No, please follow me, snowman.
S2:	Snowman.
T:	OK, great! Any comments?

In terms of the meaning-related feedback, Ms P spent relatively more time (27.7%) evaluating the meaning conveyed by the students' responses than Ms W did (18.3%). This feedback was usually either in the form of confirmation, with the teacher repeating the student's response and/or in the form of explanation, with the teacher highlighting the information conveyed. For example, in Extract 5, the teacher received information transmitted from the group Q&A and asked further convergent questions to one of the group members to assess the group's understanding. At the same time, the teacher also checked the understanding of the whole class. By contrast, in some of the peer assessment, the meaning-related feedback was limited to holistic comments like "great", "good pronunciation" or "just so-so".

Extract 5 (Ms P, Grade 4, School A) Teacher meaning-related feedback in a group Q&A assessment activity

	[One group had finished displaying its Q&A with all group members]
T:	OK, great thank you! I have one question. xxx (S1's name), you mentioned that you have a sister. A baby sister?
S1:	Yes!
T:	What's her name?
S1:	Her name is Cindy.
T:	OK, [turning to the whole class]. Xxx, has a sister and her name is...
Ss:	Cindy!
T:	Yes, I think xxx has a happy family, right?
Ss:	Yes!
T:	Ok, great. Sit down, please.

The combined feedback focused not only on matters of linguistic form but also on the expression of the meaning inherent in the students' responses. Both teachers used this kind of feedback in about half of their assessment activities, 46.1% and 52.9% in Ms P and Ms W's lessons, respectively. Extract 6 provides an example demonstrating a teacher's combined focus on form and meaning in her feedback. It might be noted that Ms P repeated the student's response by emphasising the corrected linguistic forms and then gave further explanation. In addition, she also asked follow-up convergent questions, which were not in the target form but served for communicative purposes. However, the students' response was limited to simple yes or no answers.

Extract 6 (Ms P, Grade 4, School A) Teacher form-meaning combined feedback in a Q&A led by teacher (guessing game)

T:	So, boys and girls, let's do a guessing game. Please look at my bag. Can you guess what I have? Please ask me questions using the target sentence 'Do you have ...?'
S1:	Do you have lunchbox?
T:	OK, do I have A lunchbox? No, I don't. Because I will have my lunch at home. How about you?
S1:	No, I can't.
T:	No, you DON'T have a lunchbox, right? Are you living nearby?
S1:	Yes, I am.
T:	So, you don't need a lunchbox either! OK, good, sit down please.

To sum up, the assessment practice in two Grade 4 teachers' classes was mainly teacher-led, targeting individual response with a focus on form-meaning combined feedback. During the process of class assessment, teachers usually adopted a convergent approach by asking closed questions with a single expected answer to check learners' understanding (Hill & McNamara, 2011; Torrance & Pryor, 2001). While this approach seems in line with the principles of NECS targets, which encourage the use of language in real-life settings, the approach also, nonetheless, constrained students' development of their oral competence.

4.3.2.2 Classroom-based assessment practices in Grade 6

In this section, the results of classroom-based assessment practices by two Grade 6 EFL teachers are presented in Table 4.8. The percentage represents the proportion of total assessment activity time spent on each of the three practical dimensions.

Table 4. 8 Time and percentage of time spent on different categories by two Grade 6 teachers

	Ms L, School A		Ms B, School B	
Features	**Time**	**%**	**Time**	**%**
Agent of assessment				
Teacher	69.5 min	100%	63.5 min	79.8%
Peer	0 min	0%	16 min	20.2%
Target of assessment				
Whole class	5 min	7.2%	10.5 min	13.2%
Group	12.5 min	17.9%	7 min	8.8%
Individual student	48.5 min	68.8 %	66.5 min	78 %
Actual focus of assessment (feedback)				
Forms	8 min	11.5%	17 min	21.4%
Meanings	12.5 min	18%	9.5 min	11.9%
Combination	49 min	70.5%	53 min	66.7%

Note: % indicates the percentage of total classroom-based assessment time

Agent of assessment

As was the case in Grade 4, the two Grade 6 teachers played the role of the major assessment agent in class. Specifically, Ms L was the sole agent when making assessment, and no peer assessment was observed during class. Though Ms B invited students to engage in peer assessment (20.2%) during some group activities like textbook role-play and Q&A led by the group, she still served as the main agent during assessment (79.8%).

Target of assessment

Regarding the target of assessment, similar to the results in Grade 4, individual student response was the primary target of assessment in the two Grade 6 classes (68.8% and 78%). In contrast, whole class choral answers (7.2% and 13.2%) and group responses (17.9% and 8.8%) were less frequently assessed. It could be noted that although 35.2% of assessment activities in Ms B's class were organised at the group level (see Table 4.2 in Section 4.2.2.2), only 8.8% of assessment feedback was targeted at overall group performance, which was non-specific and primarily "person or ego referenced" in nature such as "so great" and "wonderful" (Tunstall & Gipps, 1996). Instead, teachers and peers tended to give more task-referenced feedback at the individual level.

Actual focus of assessment: feedback

Turning to the actual focus of assessment in Grade 6, the two teachers spent the majority of assessment time focusing on both the linguistic forms and meanings expressed in the students' output (70.5% in Ms

L's class and 66.7% in Ms B's). These proportions were larger than those found for Grade 4. In contrast, the time spent exclusively on either linguistic forms (11.5% and 21%) or meanings (18% and 11%) was much less.

It seems that Grade 6 teachers mainly provided combined feedback during assessment, showing a balance of actual focus on language use and linguistic form. However, when looking at the content of the combined feedback, the teachers seemed to place more emphasis on the accuracy of forms. For example, a textbook retelling assessment activity (Extract 7) aimed to assess students' narrative skills based on cued language and visual aids, the actual focus of the teacher's feedback was on the linguistic form more than the meaning; the activity also strictly followed the exact sentences in the right tense forms from the textbook. Students rarely seemed to have the chance to use their own language freely or creatively.

Extract 7 (Ms L, Grade 6) Teacher form-meaning combined feedback in a textbook retelling activity

T:	OK, now I will ask someone to retell this part of the story. Who wants to have a try? Yes, you please.
S1:	Finally, Jack found the place and give the letter to Mr Brown.
T:	Yes, very good, thank you! OK, you, please.
S2:	Finally, Jack found the place and... letter to...
T:	[points to the cued word 'give' on the blackboard] Gave.
S2:	Oh, have to letter.
T:	GAVE A LETTER TO. Tell it again, please.
S2:	Jack found to place...
T:	Found THE place.
S2:	OK, Jack found the place and gave to letter...
T:	[holds the cued visual aid an envelope] GAVE A LETTER
S2:	Oh, gave a letter to Mr Brown.
T:	OK, please do it again.
S2:	Jack found the place and gave the letter to Mr Brown.
T:	OK, good, thank you. Who else wanna try?

To sum up, in two Grade 6 classes, both teachers were the major agents of assessment when evaluating students' learning evidence, and the teachers tended to provide feedback to individual students. The actual focus of the feedback, though apparently using a combined form, placed more emphasis on linguistic accuracy.

4.4 Summary of chapter four

This chapter has addressed the first main research question by presenting the findings related to classroom-based assessment activities and their implementation. Two main strands of data were reported: the teacher questionnaire of local primary EFL teachers and classroom observations of four EFL teacher at two grade levels (i.e., Grade 4 and Grade 6) from two local state primary schools. The

chapter first described the most frequent assessment activities in Grade 4 and Grade 6 and analysed which aspects were assessed in relation to specific NECS targets (RQ1.1). Then, turning to the practical level, the chapter summarised how local EFL teachers implemented these assessment activities from three dimensions (i.e., agent, target and feedback of assessment) (RQ1.2).

Chapter 5 Data analyses and results: teachers' perspectives

5.1 Introduction

This chapter reports the research findings regarding teachers' perspectives, which correspond to the second main research question:

RQ2. How are the intended NECS being interpreted by local EFL teachers?

> RQ2.1 What are teachers' general perceptions of the curriculum implementation in aural and oral teaching?
>
> RQ2.2 To what extent do teachers believe that their students will reach the curriculum targets in listening and speaking skills by the end of Grade 4 and Grade 6?

This chapter consists of two sections, with each section addressing one of the two sub-research questions. For each section, both quantitative questionnaire and qualitative interview findings are presented, with the aim of the triangulation to see whether these two strands of data align or diverge.

5.2 Teachers' general perceptions of the NECS

This section addresses the research questions of RQ2.1, *What are teachers' general perceptions of curriculum implementation in aural and oral teaching?* to explore factors that, from the teachers' perspectives, facilitate or inhibit the implementation of the NECS. Findings from the teacher questionnaire administered to a total of 103 EFL teachers from 28 state schools of the local district are presented first before sharing the results of the interviews conducted with four EFL teachers at two grade levels. The section aims to better understand the teaching contexts of the participating schools before moving on to discuss more detailed information obtained from the teacher interviews.

It should be noted here that teaching was defined broadly in this study and that the definition of teaching included instructional-embedded assessment as well since it was hard for teachers to differentiate assessment elements from their daily teachings This methodology was also consistent with the way of identifying assessment activities during the classroom observation and the term teaching was used in the questionnaire.

5.2.1 Teachers' views on NECS implementation: results from the teacher questionnaire

This section presents the factor analysis results for section B1 in the teacher questionnaire. Teacher participants were asked to indicate their attitudes towards different aspects of NECS implementation

according a 5-point Likert scale (e.g., 1 = strongly disagree; 5 = strongly agree). The questionnaire covered 5 factors that affect curriculum implementation identified in prior studies: the teachers' level of English proficiency; available teaching resources; the influence of exams; teaching time, class size and workload; and teachers' professional development. The principal components factor analysis with varimax rotation was employed to identify Chinese EFL teachers' perceptions of these underlying factors, which could influence their NECS implementation.

Table 5.1 presents the results of the factor analysis with a four-factor solution, all the factors smaller than the cut-off value of .30 were suppressed (Field, 2009). The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO = .842, which was meritorious according to Hutcheson and Sofroniou (1999). Moreover, all KMO values for individual items were higher than .71, which were well above the minimum level of .5 (Field, 2009). The results indicated that the sample size was adequate for further factorial analysis. An initial analysis was run to obtain eigenvalues for each factor in the data. There were four factors that had eigenvalues over Kaiser's criterion of 1 and, in combination, these factors explained 72.14 % of the variance.

There were 7 items loaded onto Factor 1, including two initially designed factors: teaching resources and teachers' professional development. Therefore, Factor 1 was labelled as *external support*. Factor 2 was defined by three items that were originally intended to measure the washback effect of examinations on curriculum implementation. The factor was therefore labelled as *testing washback*. There were five items loaded onto Factor 3, which were all originally designed to measure EFL teachers' oral and aural language proficiency. Among these items, two items had double loadings: Item "I'm confident in my English oral competence" was loaded both onto Factor 1 (.615) and Factor 3 (.533); Item "I'm confident in my English listening competence" double-loaded onto Factor 1 (.546) and Factor 3 (.429). Since these items examined teachers' oral competence, they were retained with Factor 3. Therefore, Factor 3 was labelled as *teachers' oral proficiency*. The remaining three items, as originally designed to measure teachers' contextual conditions and their teaching workloads, loaded on Factor 4. Thus, it was labelled as *contextual constraints*.

Table 5. 1 Results of factor analysis for the teachers' general perceptions of curriculum implementation in aural and oral teaching

Items	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
External support				
The content of the textbook provides me detailed instructions in teaching listening and speaking.	.862			
The teaching and research activities in my school improves my teaching skills in listening and speaking.	.804			
I have other resources when teaching listening and speaking.	.721			
NECS provides me detailed learning objectives in listening and speaking.	.709			
The teachers' book provides me enough guidelines on how to teach listening and speaking and activities resources.	.587			
Suggestions from the district-level ELT coordinator contribute to the improvement of my English teaching ability.	.572			
Participating in the national in-service training could help me better understand the NECS.	.520			
Testing washback				
The local final exam influences my way of teaching listening and speaking.		.839		
The local final exam influences the content of my teaching in listening and speaking.		.811		
The junior high school entry exam influences my teaching focus in listening and speaking.		.764		
Teacher's oral proficiency				
I can use English to discuss with students.			.768	
I can use English to ask high standard questions			.714	
I can use English to give detailed feedback.			.659	
I'm confident in my English oral competence.			.533	
I'm confident in my English aural competence.			.429	
Contextual constraints				
The large class size makes it challenging for me to teach speaking and listening.				.889
I have moderate teaching workloads. (reversed one)				.836
I don't have enough time to let students practice their oral English in my class.				.771
Eigenvalues	6.704	3.704	1.524	1.052
Percentage of the totally variance accounted for	37.245	20.575	8.469	5.846

Descriptive statistics including means, standard deviation and coefficient-alpha reliabilities for these factors are illustrated in Table 5.2.

The mean score for each multi-item factor was calculated to get the overall intensity of each one. The results reflected a positive attitude in factors of *external support* (M = 4.11, SD = 0.64), *teacher oral proficiency* (M = 4.06, SD = 0.53) and *contextual constraints* (M = 3.99, SD = 0.58). The relatively homogeneous response was reflected by the small standard deviation. The lowest mean score was found in *testing washback* factor (M = 2.83, SD = 1.29) with a relatively high standard deviation suggesting a negative yet less consistent attitude toward this issue. The teacher participants tended not to agree that curriculum implementation was affected by local examinations.

In addition, internal consistency for each of the factors was high: .87 for *external support*; .83 for *teacher oral proficiency*; .78 for *contextual constraints*; and.92 for *testing washback*, indicating that teachers' responses were consistent across items within these defining factors.

Table 5. 2 Descriptive statistics and internal consistency for the factors of teachers' general perceptions of curriculum implementation

Factors	**No. of items**	**Mean (SD)**	**Coefficient-Alpha Reliability**
External support	7	4.11 (0.64)	.87
Teacher oral proficiency	5	4.06 (0.53)	.83
Contextual constraints	3	3.99 (0.58)	.78
Testing washback	3	2.83 (1.06)	.92

To sum up, the factor analysis revealed four underlying factors associated with curriculum implementation of aural and oral teaching in the local context. Most of the items loaded onto the presumed factors were drawn from the literature review. In the next section, teacher interview data is presented to see if these factors were also identified therein and if there was more detailed and in-depth information was provided to better understand the local context.

5.2.2 Teachers' views on NECS implementation: results from the teacher interviews

This section presents findings from the teacher interviews. Four emergent themes were identified from four EFL teachers' interviews: knowledge and understanding of NECS; contextual conditions; teacher factors; resources support. Each theme is discussed in a separate subsection below with exemplified extracts.

5.2.2.1 Knowledge and understanding of NECS

This theme discusses four teachers' knowledge and understanding of NECS. Within the theme, two sub-categories were generated from the commonalities of the teachers' interviews: the role of NECS and the content of NECS. The different points from the teachers' interviews are also discussed.

The role of NECS

All the teachers acknowledged the fundamental role of NECS as a teaching guideline at the macro level. They identified it as a rough reference for lesson planning and classroom teaching. As their comments illustrated below,

> As an ordinary classroom teacher, the NECS is not very important; however, I will follow its requirements when conducting teaching, for example, the requirements on sentence pattern or language skills. ... For me, it's is more like a supplementary material to consult with. (Ms P, Grade 4, School A)

> I think the NECS is comprehensive and it sets a benchmark for our teachers. We have to go for these goals rather than teaching whatever we want. (Ms W, Grade 4, School B)

> To speak in a broad way, the NECS is useful for teaching. It sets a balance for me, for example, a balance of managing teaching time. It sets a standard for me on what I need to teach and what I want to teach within one class. ... I will consult with the NECS when planning the lesson, I will see whether my teaching goals match with the curriculum targets, especially if there is any excess part. (Ms L, Grade 6, School A)

> I think the NECS might provide me a rough framework. It sets the basic requirements/goals for all our classroom teachers. (Ms B, Grade 6, School B)

In addition to this role, two of the four teachers also treated the NECS as a baseline to understand their students' cognitive levels and linguistic abilities. These two teachers realized they might understand what their students could do and which levels their students were at according to the NECS. Some quotes are presented below,

> When I was finishing my teaching in reading by the end of this term for my Grade 4 students, I found there were still around 40% or even 50% of my students couldn't master the reading skills like skimming and scanning. However, with the NECS in mind, I'm not anxious any more. Because I know those students who master these skills are ahead of the curriculum goals and it is ok for those who don't as these are not required by the NECS at their current level. (Ms P, Grade 4, School A)

> I had once tried giving my students writing lessons before. However, I gave it up for various reasons, like limited time and doubts in their capability of doing it in Grade 6. Until this year, I found there were writing requirements in the NECS at Level 2, but we teachers usually ignore them. Therefore, I've started giving writing lessons to my current students and I found they could accept it. This writing course

> also works with the Grade 6 students from other primary schools, I found they can actually write something under my guidance. (Ms L, Grade 6, School A)

However, two teachers indicated that the NECS was not practical for classroom teachers with regard to their daily teaching. Rather, as their remarks, which follow, suggest, the role of the NECS was more significant for people at the administrative and policymaking levels:

> It might be of value for the district-level ELT coordinator and researchers for its theoretical usefulness and for the school administrators who might consult with it when ordering textbooks. However, for our ordinary classroom teachers, it only serves as a guide when I'm developing the teaching plan for the whole term. But it's not practical for daily teaching. (Ms P, Grade 4, School A)

> I started to pay attention to the NECS in the past two years as I need to help with the local education bureau on developing exam papers. I need to consider whether the test items are suitable for the students in the target grade and in our local district. (Ms L, Grade 6, School A)

The content of NECS

While recognising the role of NECS as a general teaching guide, all the teachers expressed their doubts about the content of NECS due to its impracticality and lack of clarity in connection to daily teaching. The teachers indicated that the descriptions in the NECS were "too general and abstract" (Ms P, Grade 4, School A). It lacked enough detailed information to support actual teaching; teachers needed to make their own efforts to make further interpretations,

> The content of [the] NECS is too general and not clearly stated; for example, the can-do statements in the language skills targets—I don't know how to implement them in my daily teaching. It's effortful and time-consuming for the teachers to integrate the curriculum targets with the textbook. (Ms B, Grade 6, School B)

> The content of NECS needs to be digested by the teachers ourselves, and we need to make further adjustments based on our actual school situation. (Ms W, Grade 4, School B)

In addition, one teacher also questioned the process of establishing the NECS targets, expressing that the ambiguity of the content made it difficult for teachers to make accurate interpretations. As she elaborated,

> It's too abstract for our classroom teachers. I don't know how the targets were produced and set at different levels. It doesn’t specify whether the NECS was designed for most of the students or this is just the basic requirements for the low-achieving students. I don't know whether the targets were set as the upper limits or the bottom line for my students. (Ms P, Grade 4, School A)

5.2.2.2 Contextual conditions

This theme, which was also frequently mentioned in the teachers' interviews, discussed contextual conditions in relation to curriculum implementation. The findings related to this theme are centred on three categories: large class size; limited instructional time and washback effects of examinations.

Large class size

All the teachers commented on the large class sizes, which contribute to the perception of heavy workloads. Three of the four teachers mentioned that "there were around 50 to 55 students in each class" (Ms P, Grade 4, School A).The teachers also said that they normally "taught the whole grade which was constituted of four classes and altogether there were around 200 students" (Ms W, Grade 4, School B). Apart from classroom teaching, the teachers were also faced with "marking all the students' assignments including class work and homework, and sometimes supervising those low-achieving students for additional assignments" (Ms B, Grade 6. School A).

Additionally, the teachers felt that large class sizes might constrain their classroom curriculum implementation, especially with regard to achieving listening and speaking targets. Teachers felt that they could not provide instant and individualised feedback on every student's performance:

> For those NECS targets like 'student can take actions based on teacher's instruction', which requires one-on-one interaction, [the target] cannot be realized in a large class. It's difficult for me to assess whether every student could do it when facing with 50 students. (Ms P, Grade 4, School A)

Two of the four teachers also mentioned perceived difficulties in teaching large classes with students of mixed learning abilities and performances. Teachers "couldn't guarantee that all the students received equal attention" (Ms P, Grade 4, School A). The teachers claimed that they gave priority to low-achieving students who might encounter problems in the learning process to ensure that "no one [was] left behind" (Ms W, Grade 4, School B). The middle -level students, between the low- and high-achieving students, were not normally the focus of the teachers' attention, as Ms P (Grade 4, School A) further elaborated,

> I normally focus on asking questions to the low-achieving students to check whether they have mastered the content of my lessons or not. I also provide some opportunities to the high-achieving students as they are always raising their hands, and I don't want to frustrate their enthusiasm. However, the middle-level students are normally keeping silent and less [frequently] being asked. (Ms P, Grade 4, School A)

Limited instructional time

Limited instructional time was another frequently mentioned contextual issue in all teachers' interviews. In two participating schools, the subject of English was scheduled for three "class hours" (each class hour = 40 minutes) per week. This amount of instructional time was allocated by the local government and also met the basic requirement set by the NECS, which "allows at least 80 to 90 minutes in total with the frequency of three to four times a week" (Chinese Ministry of Education, 2011, p. 6). However, all the teachers felt that the mandatory instructional time dedicated to EFL teaching was insufficient, making it impossible to cover all the content in the textbook, let alone to consider achieving the NECS language targets, especially for aural and oral language skills. As Ms L remarked,

> One of the biggest challenges for me is that the time is inadequate. I only have three classes per week; however, the whole unit in the textbook contains four sections. For Grade 6 students, I basically need five to six classes, which means around two weeks, to finish a unit. However, there are 12 units in total and we have just 20 weeks in each term. If I have to consider whether my lessons have met the NECS targets, I might need more time. (Ms L, Grade 6, School A)

In addition, two teachers indicated that limited instructional time might influence the amount of input and output that occurred in class. The teachers felt that there was little time available for students to participate in communicative activates and thus to improve their oral competence.

> There are around 55 students in my class and there is only 40 min for each class. On average, each student could not be guaranteed any more than one minute to express themselves or to have one-on-one interaction. Then how can you expect more from the students' output? (Ms W, Grade 4, School B)

> Students still have little chance to practice their speaking skills during class time. Due to the limited input, most of their oral production is limited to imitation. (Ms B, Grade 6, School B)

Washback effects of the examination

Concerning the washback effects of examinations on curriculum implementation, the findings (which were based on teachers' perceptions) mainly focused on the impact of two kinds of examinations: local end-of-term examinations and junior secondary school entrance examinations. All the teachers noted that the current end-of-term examination set by the local government was low-stakes. Because the local government emphasised "alleviating students' learning burdens", there was "no requirement on the scores and students were not being classified by the scores" (Ms P, Grade 4, School A). Therefore, the teaching focus was not being influenced by this exam. However, the form of the end-of-term examination was being considered as a traditional paper-and-pencil exam that did not assess students' comprehensive use of English.

> We do not have the examination pressure as the local government encourages us to focus more on the teaching. However, the form of the test is still very traditional and [is] not assessing students' communicative competence. I think my students are still learning English for the sake of test rather than using it. (Ms W, Grade 4, School B)

In contrast, two teachers in Grade 6 mentioned the negative washback effect arising from the junior secondary school entrance examinations. Although, under China's nine-year, compulsory educational system, students are automatically admitted to nearby state junior secondary schools, the students, responding to parental pressures, still compete for entrance to local private secondary schools or some state-owned but privately-funded experimental schools; acceptance to these schools requires students to achieve higher scores on the entrance examination. Grade 6 teachers found that such examination pressure influenced their daily teaching practices by committing more effort and time to the test-related content and test-taking skills.

> Since the junior entrance examination only tests reading and writing, many parents complained that the English drama activities in my class are a waste of time and they are very keen on their children's test scores. I have to give them up and prepare my students for the test. (Ms L, Grade 6, School A)

> As an English teacher, I know listening and speaking skills are as important as reading and writing skills. However, these parents are very utilitarian as they think what is going to be tested and testing skills are the most important. (Ms B, Grade 6, School B)

5.2.2.3 Teacher factors

This theme centres on the impact of teacher factors on curriculum implementation. Based on teachers' perceptions, the findings mainly cover two aspects: teachers' English language proficiency and teachers' pedagogical knowledge and skills.

Teachers' English language proficiency

All the teachers mentioned the importance of teachers' English language proficiency, especially their listening and speaking ability, and its relation to the effectiveness and success of English teaching. The teachers were aware of the importance of authentic communication between teacher and students during teaching. However, they also noted that a perceived lack of English proficiency might makes teachers feel less confident in using the target language to teach and conduct effective communicative activities,

> I've taught [for] ten years primary English, and my vocabulary has just degraded to the primary level. The reason why I don't want to set open-ended questions and why I prefer teacher-led activities is because I'm afraid of not being able to communicate with them freely in English. (Ms P, Grade 4, School A)

> A real teacher must be an English language teacher whose teaching capacity and experience might be improved and accumulate with his/her age; however, if she is not working hard, her level of language proficiency will definitely decline with time. At that time, even if she has advanced teaching skills, she could not ensure ideal learning outcomes. (Ms L, Grade 6, School A)

Two teachers further mentioned that a high level of English language proficiency, especially in the oral aspect, could ensure students a better language learning environment with more comprehensive input, which would, in turn, lead to improved output.

> I think I have a relatively higher level of oral competence, and I make full use of this advantage to create more opportunities for my students to be exposed to the target language during my class. (Ms L, Grade 6, School A)

> There is one of my colleagues whose spoken English is good. Thus, her students follow and imitate her intonation and pronunciation naturally and correctly during the class. (Ms B, Grade 6, School B)

Teachers' pedagogical knowledge and skills

In addition to teachers' language proficiency, teachers' pedagogical knowledge and skills were also raised by the two teachers as a challenge faced during the process of curriculum implementation. These teachers indicated that, given the same contextual constraints, actual teaching practices will vary in accordance with the following factors: what the teachers believe with regard to the nature of learning; how the teachers select and use materials effectively; and how the teachers carry out classroom activities Some comments are illustrated as follows,

> Given the limited resources, I think teachers need to select the right supplementary material for their students and make use of it for different purposes. For example, with a simple video, I could use it just for listening purposes; or I could use it for imitation, or I could let students discuss it around a particular topic. By doing this, all I want [for] my students is [for them] to master various skills through a simple task, rather than [for me to] immerse them in thousands of mechanical practices. (Ms L, Grade 6, School A)

> I think group activities should not be limited just to students reading the dialogue within a group. Students could benefit more from [group activities] like using target language creatively or collaborative learning if I design and manipulate the task properly. (Ms P, Grade 4, School A)

5.2.2.4 Available resources

The final theme focuses on how variously available resources influence the curriculum implementation. There were three main forms of resources revealed from the teacher interview data: textbook, supplementary materials, and teacher training support.

Textbooks

Textbooks were the core material most frequently mentioned during all four teachers' interviews. Textbooks were recommended by the local government and set as the mandatory teaching material. Teachers' views on the use of textbooks mainly touched on the following two perceived aspects of these texts: the varied difficulty levels of the content and the textbooks' problematic logic.

All the teachers were dissatisfied with the level(s) of difficulty found in the textbooks. For those teachers teaching lower grade levels (i.e., Grade 4), they felt that the content was too simple and less informative; in other words, the textbook was "no longer applicable for the current students concerning their cognitive and knowledge acceptance level" (Ms P, Grade 4, School A).

> The content of [the] textbook is very easy and limited, which is not enough for my current Grade 4 students. There are very few vocabularies to teach and many students already knew how to read them before my lesson with the help of [the] dictionary or their parents... And the text or dialogue is very simple and less attractive. (Ms W, Grade 4, School B)

In contrast, those teachers in upper grade levels (i.e., Grade 6) complained about the textbooks' being too difficult and stuffed with too much content to be taught effectively. The textbook "did not reflect a gradual process of learning", as Ms L from Grade 6 further elaborated,

> The content of [the] textbooks in Grade 3 to Grade 5 is very easy; however, there seems [to be] a sudden change in Grade 6. The textbook is becoming very difficult for my students, and there are too many grammatical points to teach. I don't think my students could master them all just within the last year of their primary education. (Ms L, Grade 6, School A)

Moreover, two teachers also perceived problems with the arrangement and setting of topics in the textbooks. Specifically, the teachers thought there were no obvious topic connections between the lessons within each unit and that this lack of coherence could lead to confusion for teachers devising lesson plans.

> For example, in the first lesson of the unit, it mainly talks about stationery, however, in the following lessons, the speaking activities and reading text show no relation with this topic. Sometimes I have to find other related lessons from other units in the textbook during the lesson planning, which is confusing, and time-consuming. (Ms P, Grade 4, School A)

> The topic setting in the textbook is not strong enough. The sentence patterns and grammatical knowledge were provided in an isolated way; they didn't set contextually. I always think that English teaching should be topic-based. Students should know how to use the language on purpose instead of purely learning the grammatical points. (Ms L, Grade 6, School A)

Supplementary materials

The two frequently mentioned supplementary materials were teaching materials provided by the local government and the textbook-related teacher's guidebook. All the teachers acknowledged that they had access to some supplementary resources, which were supported by the local educational bureau. However, they also indicated that these materials could not be referenced for textbook teaching.

> At the beginning of each term, the local government will provide us download links to some authentic teaching resources, like foreign cartoon videos and classic fairy tale recordings. However, it's not directly related with the content or topic of [the] textbook. It could only be treated as supplementary materials. (Ms B, Grade 6, School B)

This finding also reflected that the teachers felt they relied very much on the textbook and noted that they lacked some autonomy in selecting materials. As for the related teacher's guidebook, it was criticized as being out of date and of less value for teaching practices and curriculum interpretation. As the teachers commented,

> It is totally of no use for my current teaching. The content has not been changed or revised for almost ten years except the cover... The guidance on pedagogical practices is already out of date. (Ms P, Grade 4, School A)

> I think the teacher's guidebook is not practical; there are no related activity ideas or tasks available. I use it only as a reference to check the teaching objectives for each lesson during teaching planning. (Ms B, Grade 6, School B)

> It doesn't provide any interpretations or guidance on helping us connect the curriculum requirements with the textbook. (Ms W, Grade 4, School B)

Teacher training support

Teacher training support referred here to the training opportunities that existed either within school, at the municipal level and/or at the national level. All the teachers acknowledged the benefits of in-school and local government trainings, which assisted them to better understand the curriculum and to refine their teaching planning. The main forms of such trainings included ordinary classroom observation; attending demonstration lessons and teaching and research group meetings. The teachers offered the following as illustrative examples of available forms of support:

> During the teaching and research group meetings, we will communicate with each other regarding the teaching process for a certain type of lesson, for example, a reading lesson. We will settle on a general process or framework which all the teachers are suggested to follow, and teachers are also encouraged to add their own ideas during textbook implementation. (Ms W, Grade 4, School B)

> For example, during this term, the *jiaoyanyuan* (i.e., district-level ELT coordinator) has provided us [with] guidance on teaching listening and speaking after the

> classroom observations. She gave us a rough teaching design on it. Our teachers need to manipulate and adjust it by ourselves. I think this is more like a reference or a direction, but it's practical. (Ms L, Grade 6, School A)

It could also be noted that the guidance from such trainings was provided at the macro level. It still largely depended on how teachers implemented the trainings based on their understanding. In contrast, the national in-service training was viewed as less effective and impractical. Since this was a compulsory online training delivered by some university lecturers, all the teachers complained that its "formality overweighed…its content "(Ms B, Grade 6, School B). There was an apparent disconnection between the content of the teaching and the teachers' real need.

> You have to attend all the online sessions and have to answer the pop-up questions every 15 minutes to ensure that you are actually concentrating on it. However, the content is very boring. These so-called experts were not front-line or experienced teachers; they are just talking about theory, which is hard to connect with my classroom teaching. (Ms W, Grade 4, School B)

> What these training experts said was very abstract and not practical. They did not totally understand the context we are faced with. What we need is to observe some demonstration lessons by actual experienced teachers and afterward have a sort of self-reflection. (Ms L, Grade 6, School A)

5.2.2.5 Summary

This section has presented the teachers' interview findings regarding their perceptions of curriculum implementation. The results revealed four main themes that can be further categorised as internal (e.g., teacher factors) and external factors (e.g., the content of NECS; contextual conditions; available resources).

For the external factors, the following conclusions might be drawn: 1) The NECS was recognised as a guideline at the macro level for its general principals; however, teachers felt that it was written in an ambiguous and general way that confused them and left them searching for further interpretations and for more immediate connections to their daily teaching. 2) As for the contextual conditions, large classes with students at different ability levels and the limited instructional time available were seen as challenges that impeded the teachers' efforts to ensure their students received enough language exposure; these external factors also stymied the teachers' efforts to involve more interactive activities and to guarantee that every student participated in these activities and received feedback. Although teachers did not face exam pressures from the local government, teachers in the upper grade levels were influenced by the negative washback effects from the junior entrance examinations. 3) Regarding available resources, the textbook was criticized for its out-of-date content and its inconsistent levels of difficulty. In addition, the training support offered at the national level was criticized as not being valuable, although training provided by the local government was seen as beneficial.

As for the internal or teacher factors, these mainly included teachers' language proficiency and their pedagogical knowledge and skills. Teachers admitted that their level of English language proficiency influenced the quality of classroom teaching and effective communication and especially affected their ability to help students improve their oral and aural competence. In addition, teachers also indicated that, given the external factors, how they were able to minimize the influence of contextual constraints and make full use of limited resources depended on their professional knowledge and skills. This finding also suggests that there was a certain amount of interplay between external and internal factors.

5.2.3 Summary

To sum up, both the factors identified from the quantitative questionnaire results and the themes derived from the qualitative interview findings showed a certain degree of overlap as well as a divergence regarding teachers' general perceptions towards curriculum implementation. Table 5.3 summarizes the major findings from both strands of data.

Table 5. 3 Summary of factors influencing teachers' perceptions of curriculum implementation from the teacher interview and questionnaire

	Questionnaire	**Interview**
External factors	External support	The NECS
		The role of the NECS
		The content of the NECS
		Available resources
		Textbook
		Supplementary materials
		Teacher training support
	Contextual constraints	Contextual conditions
		Large class size
		Limited instructional time
	Testing washback	Washback effects of exams
Internal factors	Teacher oral proficiency	Teacher factors
		Teachers' language proficiency
		Teachers' pedagogical knowledge and skills

As for the external factors, the questionnaire data confirmed that external support, contextual constraints and testing washback were felt by teachers to affect their curriculum implementation. Among these factors, scores on external support achieved a mean of 4.11 on a scale of 5, indicating that teacher respondents in the questionnaire agreed that the current forms of external support, including the NECS, the textbook and teacher training support, facilitated their English teaching. However, while recognising the role of the NECS as well as support from the local government, the qualitative interview results also indicated that teachers felt that the content of the NECS and the textbook are weak and that the training support at the national level is impractical. With regard to contextual teaching conditions, the questionnaire data also confirmed that the large class size and limited instructional time were reported to be major constraints that impeded teachers' curriculum implementation, especially for oral skills teaching. Regarding the impact of exams on curriculum implementation, the questionnaire data, with a mean score of 2.83, revealed similar findings as the interview data, in which teachers reported that their teaching practices tended to be less influenced by exams. However, teacher interviews further indicated that teachers in the upper grade levels reported being influenced by the junior entrance exams.

5.3 Teachers' attitudes towards the NECS targets in listening and speaking

This section presents the findings for the research question RQ2.2: *To what extent do teachers believe that their students will reach the curriculum targets for listening and speaking skills by the end of Grade*

4 and Grade 6? Findings from the teachers' interviews as well as from the teacher questionnaire are presented in the following sections.

5.3.1 Results from the teacher questionnaire

This section presents the results of the teacher questionnaire, aiming to obtain a relatively comprehensive picture regarding the teachers' estimation of their students' ability to meet NECS targets at two levels in the local context.

On a 5-point Likert scale (e.g. 1 = none of my students could achieve this target; 5 = 100% of my students could achieve this target), teacher participants were asked to evaluate the extent to which their current students' English proficiency was likely to reach by Grade 4 or Grade 6 the appropriate grade-level NECS targets for listening and speaking. Those whose current students were in Grade 3 and/or Grade 4 were asked to rate based on Level 1 targets, while those who were currently teaching in Grade 5 and/or Grade 6 were asked to rate based on Level 2 targets. In the main study, there were 62 teachers who rated based on Level 1 and 58 teachers who rated based on Level 2. Teachers' mean responses regarding their evaluation of their current students' performance in two skills domains were calculated, rounded up and ordered. Table 5.4 and Table 5.5 summarise the results at Level 1 and Level 2, respectively.

Table 5. 4 Teachers' perceptions of NECS targets at Level 1

Skills	Targets	Mean (SD)	Rounded Mean
Listening	1. Can understand and react to simple classroom instructions.	4.06 (.827)	4
	2. Can recognize pictures or real objects based on heard vocabulary and sentences.	4.03 (.765)	4
	3. Can act in response to directions. (e.g., pointing pictures, colouring, drawing and acting).	3.95 (.749)	4
	4. Can understand simple stories with clues of pictures and actions.	3.21 (.861)	3
Speaking	1. Can greet in simple sentences.	4.42 (.731)	4
	2. Can exchange personal information (e.g., name and age).	4.32 (.713)	4
	3. Can imitate speaking based on tape	4.23 (.752)	4
	4. Can guess the meaning and figure out the words based on acting.	4.05 (.745)	4
	5. Can express simple emotions and feelings (e.g., like and dislike).	3.65 (.624)	4
	6. Can speak words and phrases based on picture and text.	3.45 (.733)	3

1 = Not appropriate
2 = Approximately 25% of primary school students in your class could achieve this target.
3 = Approximately 50% of primary school students could achieve this target.
4 = Approximately 75% of primary school students could achieve this target.
5 = Approximately 100% of primary school students could achieve this target.

Among the four targets for listening skills in Level 1, the ordered means indicated that the three targets related to understanding classroom instructions (Target 1: M = 4.06, SD = .827; Target 3: M = 3.95, SD = .749) and vocabulary (Target 2: M = 4.03, SD = .765) ranked higher than the target related to understanding text-level stories (Target 4: M = 3.21, SD = .861). The rounded means further indicated that the lower grade level teachers' estimations anticipated that the majority of their current students (75%) could achieve the first three targets while around half of the students could achieve the last one.

In terms of the six speaking skills targeted in NECS Level 1, the highest-ranked targets were those related to using basic formulaic exchanges, such as greetings and daily expressions (Target 1: M = 4.42, SD = .731; Target 2: M = 4.32, SD = .713). Next were the targets related to daily teaching activities, such as imitation (Target 3: M = 4.23, SD = .752) and guessing games (Target 4: M = 4.05, SD = .745). The lowest-ranked targets involved the ability to produce expressive (Target 5: M = 3.65, SD = .624) and narrative (Target 6: M = 3.45, SD = .733) phrases and sentences. The rounded means further indicated that, based on teachers' estimation, around 75% of the students could achieve most of the targets in listening skills, but only 50% of the students could achieve the target requiring narrative skills.

Table 5. 5 Teachers' perception of NECS targets at Level 2

Skills	Targets	Mean (SD)	Rounded Mean
Listening	1. Can understand common instructions and directions and respond properly.	4.16 (.664)	4
	2. Can understand simple questions in class	4.03 (.669)	4
	3. Can listen to and understand conversations and recordings with clues of pictures, images, gesture.	4.00 (.671)	4
	4. Can listen and understand picture stories.	3.37 (.691)	3
Speaking	1. Can use daily expressions (e.g. greetings, express one's thanks, apologize) and express simple emotions (e.g. like and dislike).	4.26 (.709)	4
	2. Can have simple conversation based on one's familiar personal or family situation.	3.69 (.649)	4
	3. Can narrate based on daily topics	3.62 (.639)	4
	4. Can articulate English clearly with the appropriate intonation.	3.52 (.623)	4
	5. Can describe or narrate simple stories with the help of teachers and clues of pictures.	3.41 (.786)	3

1 = Not appropriate
2 = Approximately 25% of primary school students in your class could achieve this target.
3 = Approximately 50% of primary school students could achieve this target.
4 = Approximately 75% of primary school students could achieve this target.
5 = Approximately 100% of primary school students could achieve this target.

For the four listening skills targeted in Level 2, there were three targets that were above the mean score of 4, indicating that the teachers in the upper grade levels estimated that most of the students (75%) could achieve them: these included understanding classroom instructions (Target 1: M = 4.16, SD = .664), teachers' questions (Target 2: M = 4.03, SD = .669) and basic dialogues (Target 3: M = 4.00, SD = .671) by the end of Grade 6. All of these targets were classroom-based in context and were closely related to daily teaching activities; in comparison, teachers estimated that only half of the students might achieve the listening skills target, which involved text-level listening comprehension (Target 4: M = 3.37, SD = .691) and was less frequently used during the class.

With regard to the five speaking skills targets in Level 2, the top three targets were the ones related to using daily expressions (Target 1: M = 4.26, SD = .709), holding conversations (Target 2: M = 3.69, SD = .649), and telling personal stories (Target 3: M = 3.62, SD = .639). The target for speaking English with the correct pronunciation and intonation (Target 4: M = 3.52, SD = .623) followed. The rounded means for these four targets indicated that around 75% of the students were estimated to achieve them. Similar to the results for Level 1 speaking skills targets, the target that required narrative skills was ranked the lowest, with an estimation that around 50% of students could achieve it.

5.3.2 Results from the teacher interviews

This section presents the findings from the teacher interviews. During the interviews, four teachers were asked to comment on the NECS targets and to estimate the proportion of their students that might achieve these targets by Grade 4 and Grade 6. Two emergent themes were identified from the teacher interviews: the difficulty level and the suitability of the NECS targets. These themes are discussed in relation to the NECS target level (i.e., Level 1 and Level 2) in the following sections.

5.3.2.1 Teachers' views on the NECS targets in Level 1

Both teachers from Grade 4 commented that the general difficulty level for the listening and speaking targets in Level 1 was relatively easy for their current students. The teachers stated that they did not completely follow those targets during their teaching; rather, they adjusted the targets based on their students' actual learning abilities. Given this, the two teachers from the lower grade level thought that the majority of their students could achieve, and some students could even exceed, the NECS targets by Grade 4.

> The NECS targets are a bit easy and simple for them. For example, as set in the Level 1 targets, students can take actions based on instructions like acting or colouring pictures. Actually, this was completely done in their Grade 3... Almost 70% to 80% students could achieve these [targets] by the end of Grade 4. (Ms W, Grade 4, School B)

> I think most of the teachers like me don't purely follow the national targets. I still need to consider their acceptance level and set higher or lower targets accordingly. For example, during my class, I've already trained my Grade 4 students to narrate based on given keywords, which probably belongs to the targets set for the junior secondary school students. It works fine. (Ms P, Grade 4, School A)

5.3.2.2 Teachers' views on the NECS targets in Level 2

With regard to the targets in Level 2, two upper grade level teachers thought that the general difficulty level of the targets was reasonable and that most of their middle-level students could achieve these targets by the end of Grade 6. However, given the contextual constraints discussed in section 5.2.2.2, the targets related to receptive skills (i.e., reading and listening) were perceived to be more achievable than the targets related to productive skills (i.e., speaking and writing). As Ms B (Grade 6, School B) further elaborated,

> I think over half of my students could achieve all the NECS targets in Level 2. Specifically, the targets in reading and listening were relatively easy to achieve; however, it is challenging for those low-achieving to middle-level students to achieve the speaking targets. The student's output was still stopped at the sentence level rather than the text level. (Ms B, Grade 6, School B)

5.3.3 Summary of chapter five

Drawing from both teacher interviews and teacher questionnaire data, this chapter has presented findings regarding teachers' perceptions of the NECS targets. The overall findings revealed that the majority of students were estimated to achieve most of the corresponding NECS targets in listening and speaking skills by Grade 4 and Grade 6. These targets were relatively easy to achieve as they were commonly found in everyday classroom interactions involved with formulaic and simple sentences and were closely related to students' daily life. In contrast, the targets related with text-level listening comprehension and narrative descriptions were relatively hard to achieve, with only about half of the students being expected to reach them. Moreover, this finding was also confirmed in the teacher interviews.

Chapter 6 Data Analyses and results: student outcomes

6.1 Introduction

This chapter reports research findings on students' outcomes, which correspond to the third and fourth main research questions:

RQ3. According to the targets set by the NECS, what can young EFL learners actually do in speaking and listening upon finishing Grade 4 and Grade 6 respectively?

> RQ3.1 What are Grade 4 and Grade 6 EFL students' performance on two distinct assessments, the BAS-II verbal comprehension and the Frog storytelling task?
>
> RQ3.2 How do a group of teachers interpret the Grade 4 and Grade 6 students' results on the two assessments in relation to the levels (i.e. Level 1 and Level 2) in the NECS?

RQ4. What are the attitudes of young EFL learners towards the learning, teaching and assessment of EFL in Chinese primary state schools?

> RQ4.1 What are the effective attitudes of young EFL learners at two key grade levels (i.e. Grade 4 and Grade 6) towards the EFL learning?
>
> RQ4.2 What are young EFL learners' preferences for classroom-based aural and oral assessment activities at two key grade levels (i.e. Grade 4 and Grade 6)?

It first presents the findings of the Chinese Grade 4 and Grade 6 EFL students' individual assessment results, which were mapped onto the NECS (2011 version) via the standard-setting methodology (the Angoff approach and the examinee paper selection approach, as discussed in chapter 3. Finally, it presents the results from a survey and student focus group interviews of students' affective attitudes towards EFL learning and students' views on classroom-based aural and oral assessment activities.

6.2 Young EFL learners' English oral and aural language proficiency at primary level

This section addressed the two sub research questions of RQ3 to (1) describe the research findings of two individual assessments (i.e. BAS-II Verbal comprehension and frog storytelling) and to (2) describe the descriptive results of these two assessments and then links the test results with the targets that set on the NECS via the interpretation of a panel of local expert teachers.

6.2.1 EFL students' performance results on two distinct assessments

This section addressed the research question of RQ3.1 *What are Grade 4 and Grade 6 EFL students' performance on two distinct assessments, the BAS-II verbal comprehension and the Frog storytelling task?* It presented EFL students' performance results on these two assessments. In the current study, 106 Grade 4 students and 103 Grade 6 students were individually assessed by the researcher. The BAS-II Verbal Comprehension was used to assess students' aural comprehension and the Frog storytelling task was used to assess students' oral skills. Demographic information on their chronological age and gender were detailed in Chapter 3 Section 3.3.3.2. Descriptive statistics of Grade 4 and Grade 6 students' performance on the two assessment tasks are summarised in Table 6.1 and Table 6.2.

Table 6. 1 Descriptive statistics for Grade 4 students

Measures	**No.**	**Range**	**Mean**	**SD**	**Skew.**	**Kurt.**
BAS-II						
Verbal Comprehension	106	0-22	7.94	4.53	0.87	0.23
Storytelling task						
Macrostructure						
INC score	106	0-17	5.93	3.38	0.79	0.44
Microstructure						
MLU	106	0-7.67	4.63	1.19	-0.92	3.62
TNW	106	0-252	97.61	44.26	0.23	0.89
TNDW	106	0-69	30.18	14.24	0.19	-0.14

Note: INC=Index of Narrative Complexity; MLU=Mean Length of Utterance; TNW=Total Number of Words; TNDW=Total Number of Different Words

Table 6. 2 Descriptive statistics for Grade 6 students

Measures	**No.**	**Range**	**Mean**	**SD**	**Skew.**	**Kurt.**
BAS-II						
Verbal Comprehension	103	4-29	16.17	5.78	-0.64	-0.21
Storytelling Task						
Macrostructure						
INC score	103	1-17	8.49	3.92	-0.82	-0.08
Microstructure						
MLU	103	2.5-16.64	6.77	2.06	1.61	2.07
TNW	103	6-361	149.36	61.56	0.29	0.91
TNDW	103	7-90	48.97	17.76	-0.13	-0.22

Note: INC=Index of Narrative Complexity; MLU=Mean Length of Utterance; TNW=Total Number of Words; TNDW=Total Number of Different Words

6.2.1.1 Students performance results in BAS-II Verbal Comprehension

As shown in Table 6.1, the average BAS-II Verbal Comprehension score of Grade 4 students was 7.94 (SD = 4.53) with the range between 0 and 22. For Grade 6 students (Table 6.2), the average score of the BAS-II Verbal Comprehension was 16.17 (SD = 5.78) with the range between 4 and 29. The results

of an independent sample t-test revealed that Grade 6 students scored significantly higher than the Grade 4 students on the BAS-II Verbal Comprehension, t (207) = –11.096, P<.001, *d* = 1.53.

6.2.1.2 Students performance results in Frog storytelling test

Macrostructure level

For the storytelling task, at the macrostructure level, the average Index of Narrative Complexity (INC) score of Grade 4 students was 5.93 (SD = 3.38). After 4 years of EFL learning at school, Grade 6 students showed significantly higher INC scores (Mean = 8.49, SD = 3.92), t (207) = –5.014, P<.001, *d* = 0.7.

The INC scoring scheme evaluates the main structural and story grammar elements involved in narratives. To identify trends in the elements included in the Grade 4 and Grade 6 students' language, the mean ranked lists for both grade levels were plotted (Table 6.3). The top 5 most frequent story grammar elements in the Grade 4 students' narratives were: *action, character, initiating event, setting and consequence*. These elements provided the basic structure of the story (Stein & Glenn, 1979). The top 5 most frequent story grammar elements in Grade 6 students' narratives included 4 shared with their Grade 4 counterparts, *action, initiating event, consequence* and *character*. One of the top 5 most frequent elements, *complication*, was only found in Grade 6 narratives. In the INC scoring scheme, complication is defined as "an event that prohibits the execution of a plan or action taken in response to an initiating event"(D. B. Petersen et al., 2008). In Grade 6 narratives, complexity of language was emerging.

Table 6. 3 The mean ranked list of narrative elements scores

	Grade 4	**Mean (SD)**	**Grade 6**	**Mean (SD)**
1	Action	1.05 (0.54)	Action	1.35 (0.05)
2	Character	1 (0.19)	Initiating event	1.34 (0.67)
3	Initiating event	0.89 (0.57)	Consequence	1.28 (0.77)
4	Setting	0.74 (0.61)	Character	1.03 (0.17)
5	Consequence	0.68 (0.54)	Complication	0.92 (0.06)
6	Plan	0.57 (0.27)	Internal response	0.91 (0.73)
7	Complication	0.48 (0.61)	Setting	0.87 (0.46)
8	Knowledge of dialogue	0.42 (0.59)	Causal adverbial clauses	0.42 (0.61)
9	Internal response	0.26 (0.48)	Temporal markers	0.25 (0.49)
10	Causal adverbial clauses	0.26 (0.51)	Knowledge of dialogue	0.18 (0.43)
11	Temporal markers	0.1 (0.31)	Plan	0.14 (0.42)
12	Narrator evaluations	0.02 (0.16)	Formulaic markers	0.11 (0.31)
13	Formulaic markers	0.01 (0.14)	Narrator evaluations	0.02 (0.14)

In addition, to see if there were incremental changes on these elements between Grade 4 and Grade 6 students' narrative samples, an independent sample t-test was conducted with the mean scores of the 13 narrative elements. As shown in Table 6.4, significant change was identified in 9 elements, including the complexity of episode structure (i.e. *initiating event, internal response, action, complication, consequence, knowledge of dialogue*) and linguistic components (i.e. *formulaic markers, temporal markers, causal adverbial clauses*). Among which, except the mean score of *knowledge of dialogue* was found significant decrease, the remaining 8 elements mean scores were found significant growth in Grade 6 students' narrative samples. This might be explained that Grade 6 students tended to use more descriptive statements in their narratives instead of using dialogues between characters directly.

Table 6. 4 Independent samples t-test of students INC scores of narrative elements by grade

	Grade 4		Grade 6				
	Mean	SD	Mean	SD	t-test	df	Sig (2-tailed)
Character	1	0.19	1.03	0.17	-1.15	207	.25
Setting	0.74	0.61	0.87	0.46	-1.85	207	.065
Initiating event	0.89	0.57	1.34	0.67	-5.71	207	.001*
Internal response	0.26	0.48	0.91	0.73	-7.44	207	.001*
Plan	0.57	0.27	0.14	0.42	-1.62	207	.106
Action	1.05	0.54	1.35	0.05	-3.81	207	.001*
Complication	0.48	0.61	0.92	0.06	-4.3	207	.001*
Consequence	0.68	0.54	1.28	0.77	-6.53	207	.001*
Formulaic markers	0.01	0.14	0.11	0.31	-2.66	207	.008*
Temporal markers	0.1	0.31	0.25	0.49	-2.6	207	.01*
Causal adverbial clauses	0.26	0.51	0.42	0.61	-1.99	207	.047*
Knowledge of dialogue	0.42	0.59	0.18	0.43	3.17	207	.002*
Narrator evaluations	0.02	0.16	0.02	0.14	0.42	207	.675

Microstructure

In the 3 measures taken at the microstructure level, significant differences were also found between Grade 4 and Grade 6. The average total number of words in Grade 6 students' narrative production was 149 words (Mean = 149.36; SD = 61.56). This is higher than the Grade 4 students, whose average total number of words was 98 words (Mean = 97. 61; SD = 44.26), t (207) = –7.233, P<.001, d = 0.9. As for the syntactic (i.e. Mean Length of Utterance, MLU) and semantic (i.e. Total Number of Difference Word, TNDW) complexity, Grade 6 students (Mean = 6.77; SD = 2.06) had a longer mean length of utterance than Grade 4 students (Mean = 4.63; SD = 1.19), t (207) = -9.421, P<.001, d = 1.3. The total number of different words was higher among Grade 6 students (Mean = 149.36; SD = 61.56) than Grade 4 students (Mean = 48.97; SD = 17.76), t (207) = -8.771, P<.001, d = 2.28.

6.2.2 Linking individual assessment scores with the NECS

This section addresses the second sub question RQ3.2: *How do a group of teachers interpret Grade 4 and Grade 6 students' results in relation to the levels (i.e., Level 1 and Level 2) in the NECS?* Two standard-setting methods were applied during the linking process: the modified Angoff method and the contrasting group method. During the standard-setting procedure, there were 2 main stages: familiarisation and standard setting. The results of 2 stages are reported in this section.

6.2.2.1 Standard-setting results for the listening test

Results of the familiarisation stage

During the familiarisation stage, panelists were asked to review the NECS targets and discuss what was meant by a just-qualified student on a particular NECS level. Based on the NECS's general target statements, as well as their teaching experience, they were encouraged to develop a set of agreed upon definitions for listening skills at Level 1 and Level 2 as a panel. These definitions served as a reference for the later standard-setting stage, in which panelists make judgements on the test items based on these definitions. Table 6.5 and 6.6 summarise the agreed descriptors for Level 1 and Level 2 in listening skills in comparison with the general descriptors in NECS. It is evident from the table that the panelists adjusted the original descriptors and added more details based on their familiarity with the students in local context.

Table 6. 5 General descriptors and panelists' descriptors of NECS Level 1 listening skills

General descriptors for Level 1	**Panelists' descriptors for Level 1**
1. Can recognise pictures or real objects based on verbal instruction.	1. Can recognise simple pictures or real objects based on verbal instruction that uses learned vocabulary and basic phrases. The vocabulary and sentence patterns come from the textbook or the wordlist in the NECS.
2. Can understand and react to simple classroom instructions.	2. Can understand and react to simple and familiar classroom instructions which are used repetitively by teachers in the classroom.
3. Can take action based on directions (e.g., pointing at pictures, colouring, drawing and acting).	3. Can understand short and simple conversations closely related to students' daily life (e.g., family and self-introduction) and perform routine tasks in the classroom.
4. Can understand simple stories with clues from pictures and actions.	4. Can understand simple stories with clues from pictures and actions. The language of the stories is limited, with simple and formulaic sentence patterns that recur throughout the story.

Table 6. 6 General descriptors and panelists' descriptors of NECS Level 2 listening skills

General descriptors for Level 2	**Panel Agreed descriptors for Level 2**
1. Can listen and understand conversation and recordings with clues from pictures, images and gestures.	1. Can listen and understand simple conversations and descriptions with clues from pictures, images and gestures. Can understand the main points of simple recordings, though repetition is needed sometimes.
2. Can listen and understand picture stories.	2. Can listen and understand simple picture stories when the pictures are provided ahead, and enough time are allowed to comprehend the whole set of pictures.
3. Can understand simple questions in class.	3. Can understand simple questions on limited topics taken from the textbook and daily life.
4. Can understand common instructions and directions and respond properly.	4. Can understand and respond properly to familiar instructions from their own EFL teachers in class.

The results of standard setting on BAS-II Verbal Comprehension

During the standard setting procedure, panelists were asked to estimate the probability that a just-qualified student at a particular level would answer each item correctly. Panelists referred to their agreed-upon descriptors of NECS Levels 1 and 2 in making the judgements. There were two rounds of judgement, each of which was followed by a group discussion.

Panelists recommended cut-off scores on the BAS-II Verbal Comprehension for Level 1 and Level 2 in each round, which are presented in Table 6.7 and Table 6.8. The recommended cut-off scores for each panelist were calculated by multiplying the averaged estimated probabilities across items by the total scores of the test. The mean judgement for each round, its associated statistics (median, standard deviation, maximum and minimum) and the t-test results for both rounds are also provided at the bottom of the tables. The standard deviation indicates variability between panelists' judgements. From both tables, it is evident that, during Round 1, the variability (i.e., standard deviation) was higher (round comparison in Level 1: $t(22) = 3.136$, $P<.01$, $d = 1.13$; round comparison in Level 2: $t(22) = 2.25$, $P<.05$, $d = 0.96$). However, after a panel discussion on the Round 1 results, convergence was observed in Round 2. The cut-off scores for the BAS-II Verbal Comprehension at Level 1 and Level 2 were obtained from the mean cut-off scores in Round 2: 9.44 and 18.01. They were rounded up to the nearest whole number to arrive at final cut-off scores of 9 for Level 1 and 18 for Level 2. In other words, those Grade 4 students who scored 9 or above would be regarded as achieving Level 1. In Grade 6, those students who scored 18 or above would be regarded as achieving Level 2.

Table 6. 7 Panelists' cut-off score judgements on BAS-II Verbal Comprehension, Level 1

Panelist ID	**Round 1**	**Round 2**
Panelist 1	8.65	8.5
Panelist 2	10.4	9.55
Panelist 3	10.55	9.60
Panelist 4	10.55	9.85
Panelist 5	9.25	8.65
Panelist 6	12.7	10
Panelist 7	9.95	9.35
Panelist 8	10.6	9.25
Panelist 9	11.15	9.55
Panelist 10	10.1	9.55
Panelist 11	12.1	9.8
Panelist 12	10.2	9.60
Mean	10.39	9.44
Median	10.55	9.55
SD	1.1	0.45
Maximum	12.7	10
Minimum	8.65	8.5
t-test on mean	3.136	
df	22	
Sig (2-tailed)	.005	

Table 6. 8 Panelists' cut-off score judgements on BAS-II Verbal Comprehension, Level 2

Panelist ID	Round 1	Round 2
Panelist 1	19.1	18.10
Panelist 2	18.75	17.35
Panelist 3	19.39	18.30
Panelist 4	20.1	18.65
Panelist 5	18.05	17.45
Panelist 6	20.7	19.10
Panelist 7	17.15	17.05
Panelist 8	18.35	17.55
Panelist 9	20.9	19.15
Panelist 10	17.35	16.30
Panelist 11	21.15	19.70
Panelist 12	18.35	17.45
Mean	19.15	18.01
Median	18.92	17.82
SD	1.35	0.99
Maximum	21.15	19.7
Minimum	17.15	17.35
t-test on mean	2.25	
df	22	
Sig (2-tailed)	0.034	

Based on panelists' final recommended cut-off scores for Level 1 and Level 2, Table 6.9 and Table 6.10 present the number and percentage of Grade 4 and Grade 6 students by target level in listening skills. The results of Grade 4 students showed that, based on students' performance on BAS-II Verbal Comprehension, a relatively small proportion of Grade 4 students (38.68%) achieved Level 1 while a large proportion of students (61.32%) were judged below Level 1. The independent sample t-test revealed a significant difference between both groups, t (105) = -15.26, P<.001, d = 2.82.

Less than half of the Grade 6 students (45.63%) were judged at Level 2 or above and 54.37% students were judged not to have achieved Level 2 by the end of their primary education (i.e., Grade 6). The independent sample t-test indicated the achieved Grade 6 students scored significant higher than those who were below Level 2 in the listening test, t (105) = -13.31, P<.001, d = 2.94.

Table 6. 9 Distribution and t-test of Grade 4 students by listening skill level

	N	Percentage	Mean (SD)	t-test	df	Sig (2-tailed)
Below Level 1	65	61.32%	5.01(1.95)	-15.26	105	.001
Level 1 or above	41	38.68%	13.35 (3.69)			

Table 6. 10 Distribution and t-test of Grade 6 students by listening skill level

	N	Percentage	Mean (SD)	t-test	df	Sig (2-tailed)
Below Level 2	56	54.37%	11.81 (4.07)	-13.31	102	.001
Level 2 or above	47	45.63%	22.19 (2.87)			

6.2.2.2 Standard-setting results for the speaking test

In this section, the standard-setting results for the frog storytelling task was presented. The contrasting group method was applied. There were two main stages to the standard-setting procedure: familiarisation and standard setting. The results of two stages are reported in the next two sections.

The results of familiarisation

As in the familiarisation stage of the listening standard setting, panelists were asked to develop a set of specific descriptors of speaking skills for the just-qualified student at particular level based on the general descriptors in the NECS and their actual teaching context. Table 6.11 and 6.12 summarise the agreed-upon definitions for Level 1 and Level 2 in speaking skills, compared with the general descriptors in NECS.

Table 6. 11 General descriptors and panelists' descriptors of NECS Level 1 speaking skills

General descriptors for Level 1	**Panelists' descriptors for Level 1**
1. Can imitate speech based on recordings.	1. Can imitate single words, short sentences and intonation based on recordings.
2. Can greet in simple sentences. 3. Can exchange personal information (e.g., name and age). 4. Can express simple emotions and feelings (e.g., like and dislike).	2. Can speak in simple sentences and isolated phrases learned from the textbook or dealing with familiar and daily topics (e.g., exchanging personal information; expressing simple emotions).
5. Can guess the meaning of words based on acting.	3. Can guess the meaning of words based on acting. The words are limited and include some nouns learnt from the textbook and very few verbs (e.g., have and go). The lexical categories belong to daily life.
6. Can speak words, phrases and simple sentences based on pictures.	4. Can speak limited words or fixed-patterned phrases based on pictures, however, the logic of speech is not yet strong. Can pronounce monosyllabic words.

Table 6. 12 General descriptors and panelists' descriptors of NECS Level 2 speaking skills

General descriptors for Level 2	**Panelists' descriptors for Level 2**
1. Can articulate English clearly with the appropriate intonation.	1. Can articulate English clearly, however intonation and stress on some question sentences are not natural (e.g., there is a rising tone to general questions).
2. Can have simple conversations based on familiar personal or family situations.	2. Can have simple conversations on familiar daily life topics, based primarily on the simple exchange of information.
3. Can use daily expressions (e.g., greetings, gratitude expressions and apologies) and express simple emotions (e.g., like and dislike).	3. Can use daily expressions in proper situations in addition to some cultural expressions (e.g., expressions of apology and responses to gratitude).
4. Can narrate stories based on daily topics.	4. Can talk about daily life and narrate stories on daily topics using simple sentences.
5. Can narrate simple stories with the help of teachers and clues from pictures.	5. Can retell stories from textbooks with the help of teachers and clues from pictures. It is still challenging to tell simple stories if the materials are unfamiliar.

The results of standard setting on the frog storytelling test

The cut-off scores for the frog storytelling test at Level 1 and Level 2 were established using contrasting group methods. Each panelist listened to a set of oral samples and classified students into 2 categories (mastery/non-mastery) based on the degree of their oral skills with respect to the agreed descriptors for

just-qualified students at target level. Every Grade 4 student was categorised as either below Level 1 or Level 1 or above. Every Grade 6 student was classified as either lower Level 2 or Level 2 or above. Once the panelists classified the students' oral samples without knowing their actual scores, the scores earned by each student classified in a particular category were used to construct distributions and the cut score was located midway between the means of the 2 distributions (Cizek & Bunch, 2007). Figure 6.1 and Figure 6.2 illustrate the frequency distributions of Grade 4 and Grade 6 students' frog storytelling test scores in both categories based on panelists' classification.

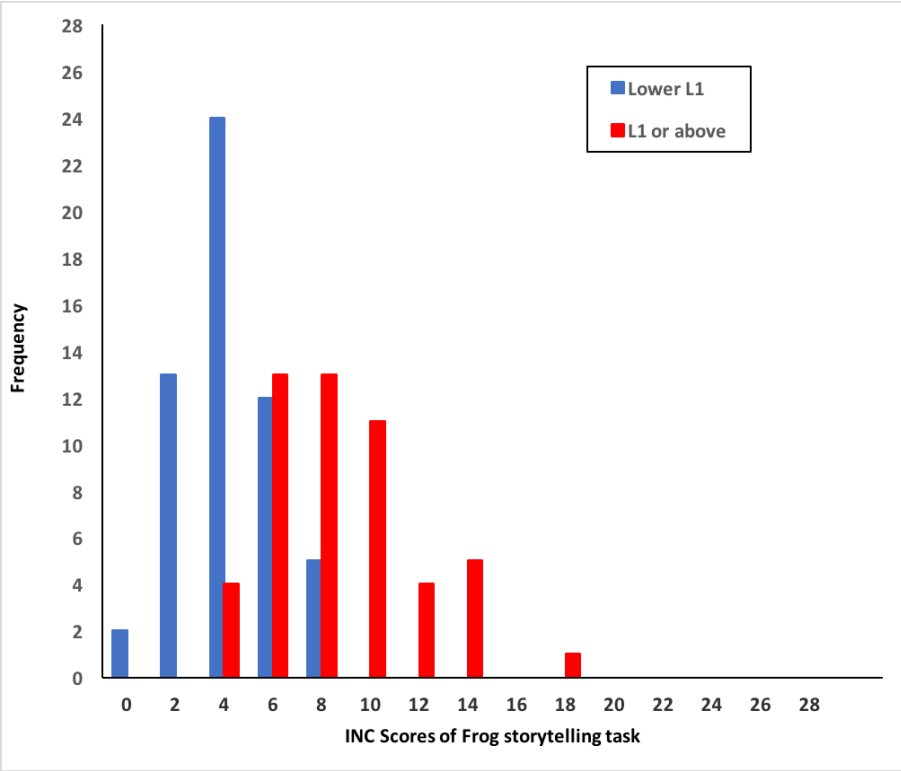


Figure 6. 1 Frequency distributions of Grade 4 students' frog storytelling test scores in 2 categories

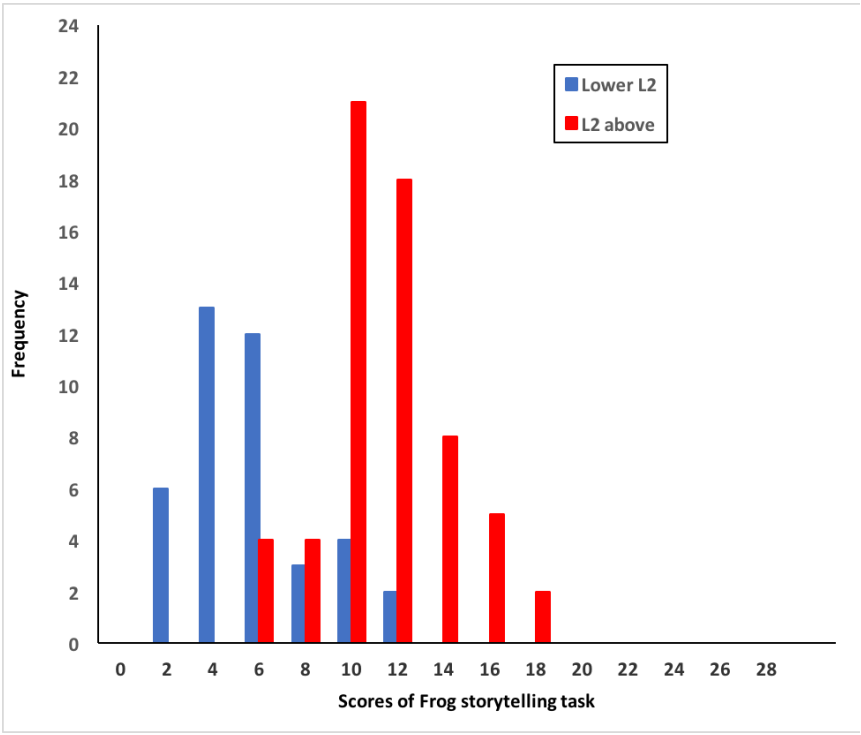


Figure 6. 2 Frequency distributions of Grade 6 students' frog storytelling test scores in 2 categories

Both figures show a certain overlap between the 2 distributions. If the cut-off score is applied midway between the means of 2 distributions, as in Figure 2, the mean score of the students classified as lower Level 1 was 3.48 (SD = 0.23) and the mean score of the students classified as Level 1 or above was 8.17 (SD = 0.34). In this case the provisional cut score would be the value midway between the 2 means, or 5.82. However, this value would need to be further adjusted to accommodate misclassification by the panelists (Kosuke Imai, 2017). There are 2 types of correct classification, true positive and true negative. Similarly, there are 2 kinds of misclassification (Cizek and Bunch, 2006). A false positive misclassification occurs when a student is not qualified at a specific level but obtains a test score above the set cut-off score. A false negative classification occurs when a student has actually mastered the skills at a specific level but obtains a test score that is below the set cut-off score. Table 6.13 illustrates a confusion matrix, also called an error matrix, where 2 types of correct classification and misclassification are shown. The misclassification rate reflects the proportion of false negatives and false positives out of the total classifications made.

Table 6. 13 Confusion/error matrix (Kosuke Imai, 2017)

	Actual outcome	
Panel classified outcome	**Positive**	**Negative**
Positive	True positive	False positive
Negative	False negative	True negative

Therefore, in order to identify a cut-off score with minimal misclassification error and make a rational choice on the final cut-off score, according to the confusion matrix, decision tables for several potential cut-off scores were constructed and the percentage of misclassifications derived from each potential cut-off score calculated. Table 6.14 and Table 6.15 presents the decision table for cut scores at Level 1 and Level 2 respectively.

Table 6. 14 Decision table for potential cut-off scores for the frog storytelling test at Level 1

If applied at:	**Cut-off = 4**		**Cut-off = 5**		**Cut-off = 6**		**Cut-off = 7**		**Cut-off = 8**	
Classified as:	**L1 -**	**L1 +**	**L1 -**	**L1 +**	**L1 -**	**L1 +**	**L1 -**	**L1 +**	**L1 -**	**L1 +**
Below cut score	24	0	38	4	47	10	50	16	54	24
Cut score or above	31	52	17	48	8	42	5	36	1	28
Total N	55	52	55	52	55	52	55	52	55	52
Misclassifications	28.97%		19.63%		**16.82%**		19.63%		23.36%	

Note: L1 - means below Level 1 and L1 + means Level 1 or above.

Table 6. 15 Decision table for potential cut-off scores for the frog storytelling test at Level 2

If applied at...	**Cut-off = 6**		**Cut-off = 7**		**Cut-off = 8**		**Cut-off = 9**		**Cut-off = 10**	
Classified as:	**L2 -**	**L2 +**	**L2 -**	**L2 +**	**L2 -**	**L2 +**	**L2 -**	**L2 +**	**L2 -**	**L2 +**
Below cut score	26	1	32	5	34	6	35	9	38	19
Cut score or above	15	62	9	58	7	57	6	54	3	44
Total N	41	63	41	63	41	63	41	63	41	63
Misclassifications	15.4%		13.5%		**12.5%**		14.4%		20.2%	

Note: L2 - means below Level 2 and L2 + means Level 2 or above.

In the above 2 tables, the percentage of misclassifications was calculated based on the proportion of misclassified students out of the total number of students. For example, in Table 3.14, if the cut score was applied at 5, 17 students who were judged as below Level 1 pass the test and then belong to the group assessed at Level 1 or above, having been misclassified due to a false positive. Meanwhile, 4 students who were judged at Level 1 or above fall below the cut-off score—a false negative misclassification. By adding these two groups of misclassified students together and calculating the misclassified proportion of the total, the misclassification percentage for the cut-off score is determined. In this instance, the misclassified percentage for the cut-off score of 5 is 19.63%.

In the 2 decision tables, it is evident that the percentage of misclassifications is varied when different cut-off scores applied. The misclassification rate reached its minimal value when the cut-off score was 6 (16.82% misclassification error) in Table 3.14 and when the cut-off score was 8 (12.5% misclassification error) in Table 3.15, which were set as the final cut-off scores for the 2 levels.

Using the final cut-off scores for Level 1 and Level 2, Table 6.16 and Table 6.17 present the number and percentage of Grade 4 and Grade 6 students that achieved the target level in speaking skills. The results showed that less than half of the Grade 4 students (46.23%) were classified as qualified students at Level 1 based on their performance on the frog storytelling task. The independent sample t-test indicates that there was a significant difference between both groups, $t(105) = -13.41$, $P<.001$, $d = 2.56$.

For the Grade 6 students, a relatively large proportion of students (61.17%) were judged as achieving Level 2 in speaking skills and their scores in the speaking test was significantly higher than the scores of those who were judged below Level 2, $t(101) = -16.78$, $P<.001$, $d = 3.48$.

Table 6. 16 Distribution of Grade 4 students by speaking skill level

	N	**Percentage**	**Mean (SD)**	**t-test**	**df**	**Sig (2-tailed)**
Below Level 1	57	53.77%	3.42 (1.43)	-13.41	105	.001
Level 1 or above	49	46.23%	8.78 (2.59)			

Table 6. 17 Distribution of Grade 6 students by speaking skill level

	N	Percentage	Mean (SD)	t-test	df	Sig (2-tailed)
Below Level 2	40	38.83%	4.3 (1.74)	-16.78	101	.001
Level 2 or above	63	61.17%	11.15 (2.17)			

6.3 Students attitudes towards EFL learning and assessment

In addressing the study's fourth main research question, *what are the attitudes of young EFL learners towards the learning, teaching and assessment of EFL in Chinese primary state schools?*, this section discusses the research findings from the student questionnaire. It first presented students' general affective attitudes towards EFL learning. Then students' view towards classroom-based assessment activities was described.

6.3.1 Students' attitudes towards EFL learning

This section presents the findings for the research question of RQ4.1: *What are the affective attitudes of young EFL learners at two key grade levels (i.e. Grade 4 and Grade 6) towards the EFL learning in a sample of local Chinese primary state schools.* In the student questionnaire, the students were asked to indicate the extent to which their attitudes towards EFL learning using a smiley face 5-point Likert-scale (e.g. ☹☹ = strongly disagree; ☺☺ = strongly agree). The questionnaire originally covered 4 types of affective attitudes: anxiety, confidence, intrinsic motivation and instrumental motivation. To explore the interrelationship among the items and to identify the main constructs of Chinese primary EFL students' affective attitudes, an exploratory principal component analysis was first carried out by including both grade levels simultaneously. And further comparisons were made between grade levels to see if there were any attitudes change among the identified constructs.

Table 6.18 presents the results of principal component analysis for the factors of students' affective attitudes in both grade levels. The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis, KMO=.88, which is well above the minimum criterion of 0.5 (Field, 2009). It indicated that the sample size was adequate for further factorial analysis.

An initial analysis was run to obtain eigenvalues for each factor in the data. Four factors had eigenvalues over Kaiser's criterion of 1 and in combination explained 50.36% of the variance. The varimax rotated factor matrix showed appreciable loadings. Seven items measuring students' anxiety in EFL learning loaded onto Factor 1, labelled *anxiety*. There were 6 items loaded onto Factor 2. These items were mainly related with students' perceptions of the pragmatic benefits of EFL learning, labelled as *instrumental motivation*. Six items related to confidence in EFL learning loaded onto Factor 3. This factor corresponded to student's confidence in learning and using English at school, thus labelled as

confidence. There were five items loaded onto Factor 4, which were related with student's enjoyment of EFL learning, labelled as *intrinsic motivation*.

Table 6. 18 Results of factor analysis for the students' affective attitudes towards EFL learning

Items	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
Anxiety				
I can feel my heart pounding when I am going to be called on in English class	.698			
I get nervous when the teacher asks me to respond to her/him in English.	.660			
I worried when I do not understand what the teacher is saying in English.	.656			
I feel badly when I cannot answer the teacher's questions in English.	.645			
I'm nervous when I hear new words or unfamiliar words in English.	.635			
I'm nervous if somebody speaks English too quickly.	.540			
I worry if I can acquire new words and expressions in English.	.379			
Instrumental motivation				
I want to be fluent in English because speaking English is useful.		.710		
I think knowing both English and Chinese helps one to get a good job in the future.		.582		
I think that English will be useful for me in the future.		.574		
I want to be good at English because this is a compulsory subject in primary education.		.546		
English is an important language to learn because it is widely spoken.		.528		
I want to get good grades in English.		.496		
Confidence				
I think that I speak English better than others in my class.			.663	
English class is easy for me.			.658	
I think I can be fluent in spoken English.			.630	
I am good at learning English.			.561	
I am confident when speaking English in class.			.397	
I am doing well in my English class.			.432	
Intrinsic motivation				
I want to know more about foreign countries.				.605
I want to make many foreign friends.				.594
It is good for Chinese to be able to speak both English and Chinese.				.532
I want to be able to speak both English and Chinese well.				.443
Learning English is one of the most important things in my life.				.305
Eigenvalues	7.143	3.855	1.882	1.222
Percentage of the totally variance accounted for	25.51	13.766	6.720	4.365

Descriptive statistics including means, standard deviation and coefficient-alpha reliabilities for each factor were illustrated in Table 6.19. The mean score for each factor was calculated to obtain its overall orientation. The results showed that a relatively strong orientation was found in two factors *instrumental* (Mean = 4.38, SD = 0.62) and *intrinsic motivation* (Mean = 4.23, SD = 0.68). It revealed that students were positively motivated to learn English for the purpose of enjoyment and for pragmatic use. In contrast, a relatively neural orientation was presented in the other two factors which were *anxiety* (Mean = 3.19, SD = 0.95) and *confidence* (Mean = 3.17, SD = 0.77). It indicated that students displayed subtle anxiety towards language learning and use at school and their confidence level were retained at middle but in a positive direction.

In addition, the reliability of each factor scale was calculated, and the results showed that *anxiety* (α = .81), *instrumental motivation* (α = .79)*, confidence* (α = .81) and *intrinsic motivation* (α = .75) all had a high level of internal consistency.

Table 6. 19 Descriptive statistics and internal consistency for the students' attitudes factor scales

Variables	**No. of items**	**Mean (SD)**	**Coefficient-Alpha Reliability**
Instrumental motivation	6	4.38 (0.62)	.79
Intrinsic motivation	5	4.23 (0.68)	.75
Anxiety	7	3.19 (0.95)	.81
Confidence	6	3.17 (0.77)	.81

An independent t test was carried out to check if there were any significant affective changes between Grade 4 and Grade 6 students. Statistically significant difference was defined at or below p = .05. As shown in Table 6.21, there was no significant differences identified in most of the affective factor scales between two groups of students except the factor *instrumental motivation*. Grade 6 student participants (Mean = 4.37; SD = 0.69) shown relatively less motivation on the pragmatic use of English than the Grade 4 student participants (Mean = 4.39; SD = 0.69), t (220) = 0.3, P =.05, d = 0.05. But the mean scores of *instrumental motivations* for each grade group were the highest among all the other scale factors and displayed positive orientation.

Table 6. 20 Independent samples t-test of students mean responses to affective factor scales by grade

	Grade level	**N**	**Mean (SD)**	**t-value**	**df**	**Sig (2-tailed)**
Instrumental motivation	Grade 4	106	4.39 (0.55)	0.33	220	0.05*
	Grade 6	116	4.37 (0.69)			
Intrinsic motivation	Grade 4	106	4.37 (0.61)	-0.73	220	0.11
	Grade 6	116	4.44 (0.76)			
Anxiety	Grade 4	106	3.24 (0.88)	0.92	220	0.79
	Grade 6	116	3.14 (0.89)			
Confidence	Grade 4	106	3.14 (0.88)	-0.62	220	0.09
	Grade 6	116	3.20 (0.85)			

6.3.2 Student preferences for classroom-based aural and oral assessment activities

This section addresses research questions RQ4.2: *What are young EFL learners' preferences for classroom-based aural and oral assessment activities at two key grade levels (Grade 4 and Grade 6)?* It aims to obtain students views and perceptions toward a series classroom-based aural and oral assessment activities. Findings were presented from two strands of data: quantitative student questionnaire and qualitative student focus group interview.

6.3.2.1 Results of student questionnaire

In the student questionnaire, students were asked to indicate the extent to which they think a list of 18 classroom-based aural and oral assessment activities would be helpful for their English learning. They were asked to rate the activities using a 5-point Likert scale (e.g. 1 = very useless; 5 = very useful). The mean preference of each assessment activity for Grade 4 and Grade 6 students was calculated and the rankings were depicted in descending order in Table 6.21.

Table 6. 21 The mean ranked list of classroom-based aural and oral assessment activities by Grade 4 and Grade 6 students

	Grade 4	**Mean (SD)**	**Grade 6**	**Mean (SD)**
1	Textbook role play	4.16 (0.91)	Q&A led by the teacher	4.29 (0.81)
2	Listening comprehension	4.15 (0.92)	Oral translation	4.29 (0.87)
3	English drama	4.14 (0.73)	Dictation	4.22 (0.96)
4	Textbook retelling	4.11 (0.88)	Listening comprehension	4.21 (0.85)
5	Oral translation	4.11 (0.89)	Listen and answer	4.18 (0.87)
6	Dictation	4.10 (1.01)	Make up conversations	4.12 (0.96)
7	Picture description	4.09 (0.59)	Storytelling	4.06 (0.83)
8	Recite text/dialogues	4.07 (0.91)	Textbook retelling	4.06 (0.91)
9	Q&A led by the teacher	4.06 (0.86)	Read aloud	3.99 (1.09)
10	Make up conversations	4.05 (0.87)	Picture description	3.98 (0.95)
11	Oral gap filling	4.03 (0.82)	Oral information gap	3.98 (0.84)
12	Read aloud	4.02 (1.10)	Make up sentence	3.98 (0.90)
13	Storytelling	3.98 (0.91)	Recite text/dialogues	3.96 (1.04)
14	Make up sentences	3.89 (0.93)	Textbook roleplay	3.94 (0.99)
15	Listen and answer	3.85 (0.94)	English drama	3.91 (0.93)
16	Listen and act	3.85 (0.96)	Listen and act	3.83 (1.01)
17	Listen and imitate	3.81 (1.06)	Listen and imitate	3.82 (1.03)
18	Q&A led by the group/pairs	3.77 (1.05)	Q&A led by the group/pairs	3.74 (0.98)

According to the table, Grade 4 students hold positive attitudes towards most of the assessment tasks. The mean ratings on 12 activities were above 4 (i.e. 4 = useful in the questionnaire). The top 5 most useful assessment activities rated by students were textbook role play, listening comprehension, English drama, textbook retelling and oral translation. Among which, textbook role play (Mean = 4.16, SD = 0.91), English drama (Mean = 4.14, SD = 0.73) and textbook retelling (Mean = 4.11, SD = 0.98) provided students more opportunities to practice their oral competence in a meaningful and interactive context. However, listening comprehension (Mean = 4.15, SD = 0.92), which was commonly found in the paper and pencil exam; and oral translation (Mean = 4.11, SD = 0.89), which was form focused, were also rated as most useful activities by Grade 4 students. In contrast, among the ranked list, the relatively less useful activities were listen and act (Mean = 3.85, SD = 0.96), listen and imitate (Mean = 3.81, SD = 1.06), and Q&A led by the group/pairs (Mean = 3.77, SD = 1.05). These activities involved less and limited oral production in compared with those top-rated ones.

For Grade 6 students, in general, around half of the listed activities were rated over 4, which was useful in the questionnaire. The top 5 most useful assessment activities were Q&A led by the teacher, oral translation, dictation, listening comprehension and listening and answer. Of those activities, oral translation (Mean = 4.29, SD = 0.87), dictation (Mean = 4.22, SD = 0.96) and listening comprehension (Mean = 4.21, SD = 0.85) were all traditional assessment activities. Q&A led by the teacher (Mean = 4.29, SD = 0.81) and listen and answer (Mean = 4.18, SD = 0.87) were all involved interactions with teacher. The relatively less useful activities rated by the Grade 6 students were listen and act (Mean = 3.83, SD = 1.01), listen and imitate (Mean = 3.82, SD = 1.03), Q&A led by the group/pairs (Mean = 3.74, SD = 0.98), which were the same as Grade 4 students.

An independent sample t-test was used to see if there were any attitudes changes among those assessment activities between two grade level students. Significant attitude changes were found on 3 activities: English drama, Q&A led by the teacher and listening and answer. Among which, English drama ($t(220) = 2.083$, $P<.05$, $d = 0.31$), which was highly preferred by younger students (i.e. Grade 4), was relatively the least-preferred activity among older students (i.e. Grade 6). Q&A led by the teacher ($t(220) = -2.106$, $P<.001$, $d = 0.29$) and listen and answer ($t(220) = -2.653$, $P<.001$, $d = 0.35$) were rated as more useful ones by Grade 6 students than Grade 4 students.

6.3.2.2 Results of student focus group interview

This section presents the findings from student focus-group interviews. There were six low achievers and six high achievers from each class at two grade levels from two schools were invited to participate. All data was gathered from eight focus group interviews. The students were asked about their views and preferences towards the classroom-based aural and oral assessment activities. In the following sections, the most preferred activities mentioned by low and high achievers in Grade 4 and Grade 6

were presented in together with their perspectives. Commonalities and differences within the interview data were summarised with multiple quotations.

Views from Grade 4 students

The most useful classroom-based aural and oral assessment activities that frequently mentioned by both Grade 4 high and low achievers were identified: Q&A led by the group/pairs, textbook retelling and listening comprehension. The first two were related to communicative and narrative skills. The last one, though involved with comprehensive listening skills, was in a traditional pencil and paper form that commonly found in the textbook exercises and examinations.

For the Q&A led by the group/pairs, both high and low achievers held positive perceptions, yet in different angles. Grade 4 high achievers regarded it as an interesting activity which involved exchanging personal information to better understand their peers and as a challenging task to get the opportunities to present their Q&A in front of the whole class. As the following quotations illustrated,

> I like it (the group led Q&A) and it's useful for practising my oral English, because we usually ask each other questions, for example, what did you do last Sunday, and I can get to know my group members better. (H3_G4SA)

> We can present our group Q&A to the whole class and everyone got a chance to say something. It's showtime for our group, how exciting! (H1_G4SB)

In contrast, low achievers found it useful from the peers' support aspect.

> I can seek help from my peers during the group Q&A activity like how to say something in English. (L2_G4SA)

> Sometimes I don't know how to read the newly learned words and our group leader will help me with this. (L3_G4SB)

For the textbook retelling activity, the high achievers found its form was challenging as they could master the ability to narrate the text by the cued language or pictures given by the teacher.

> Our teacher usually draws some pictures on the blackboard and their corresponding keywords. It's very useful to remind us of the key information of the text and we could retell it successfully! (H4_G4SA)

However, for the low achievers, the main beneficial aspect of the textbook retelling activity was to better understand the text and to memorise more key vocabularies and expressions in preparation for the examination.

> I think to retell the text is like a way of helping me to recite the text. Once I remember it, I would do it correctly when the text was appeared and tested in the final exam. (L5_G4SA)

Both high and low achievers expressed their positive orientation to the paper and pencil form of listening comprehension exercise. The following quotes reflected the washback effects of exams on students' preference on such activity.

> I would be feeling great if I got all the answers correct and I could imagine that I will also do the best during the exam. (H2_G4SA)

> I could get familiar with the exam forms and get more practices experience from these exam-liked exercises. (L6_G4SB)

In addition to the commonly mentioned useful assessment activities above, low achievers from both schools also expressed their preferences on dictation, recite and read aloud. The first two activities were memorisation based at vocabulary and discourse level. The low achievers thought the dictation would "help them to identify the vocabularies that they might be unfamiliar with by just hearing" (L2_G4SA). Recite activity, though commented as "boring" (H2_G4SB; H5_G4SA) by the high achievers, was regarded as an effective way to "review and speak out the learned words and sentences" (L4_G4SB). For reading aloud, they pointed that this would facilitate their correct pronunciation on new words. Because they feared of "being laughed at by peers when pronouncing the words wrongly" (L3_G4SA). All the three activities were form-focused and involved no creative use of the target language.

In comparison with the findings of the student questionnaire, the consistency and difference between the results of the student questionnaire and focus group interview were noted. Two activities, textbook retelling and listening comprehension, which were rated highest in the questionnaire, were also confirmed as most favoured and useful ones during the interview. However, the Q&A led by the group, though positively and frequently mentioned during the focus group interview, was ranked as the lowest one among all the activities in the student questionnaire yet with relatively greater variation, which reflected students' divergent opinions toward it and this was also revealed in the interview data in which high and low achievers held different perspectives toward this activity.

Views from Grade 6 students

For both low and high achievers in Grade 6, the frequently mentioned as most useful assessment activities were Q&A led by the teacher, make up conversation and textbook retell. The first two activities targeted at communicative skills. Students could interact with their teachers and peers. The textbook retell activity was targeted at students' narrative skills.

Compared with the Grade 4 students who favoured the Q&A within the group, Grade 6 students expressed their preference on interacting with the teacher (i.e. Q&A led by the teacher). For high achievers, they reported the enjoyment of involving genuine communications with the teacher based on their real-life experience. Exchanging unknown information and creatively using the target language was seen as proof to demonstrate their comprehensive language use ability. As the following quotes illustrated,

> I like answering teacher's questions, it's the real communication! I can share my daily life activities with Ms W. (H2_G6SB)

> Through Q&A with teacher, I can demonstrate my true English ability to the whole class. Because I'm using it! (H3_G6SA)

However, for low achievers, the beneficial aspect of the Q&A led by the teacher came from teacher's help and feedback on their pronunciation instead of the meaning expressed. Because they tended to give the expected answers from the textbook and avoided using unfamiliar words or sentences during the interaction.

> During the Q&A with Ms L, if I don't know how to answer the questions, Ms L will prompt me. (L5_G6SB)

> This is a better to practice the pronunciation of the pre-learned words. But I feared of making mistakes, so I just use the sentence patterns and words from the textbook. (L2_G6SA)

For the make-up conversation activity, high achievers favoured its practical and interactivity. It created a real context for students to use the pre-learned language to exchange new information. As they further elaborated,

> When making up the conversation, we could not only use the new sentence patterns and words from this lesson but also integrating our previous learned ones. You also need to take grammar into considerations. (H1_G6SA)

> You can create whatever a context you like, and you can use your imaginations on developing the dialogue. It could immerse everyone into a real context. It's quite fun and we can practise our oral skills. (H3_G6SB)

In contrast, the positive aspects seen from the low achievers were a sense of participation and an opportunity to review and to consolidate the pre-learned words and phrases by reciting dialogue parts assigned by their peers.

> I do not need to worry that much as I was normally assigned the relatively easier parts and I could recite it before we presenting to the whole class. But it's great that everyone could get involved! (L1_G6SA)

For the textbook retell activity, the positive perceptions between high and low achievers also displayed different angles. Grade 6 high achievers enjoyed the way of narrating the main ideas of the text using their own words based on the cued language and pictures given by the teacher. However, low achievers treated it as a good way to memorise the keywords and the whole text.

> Unlike recite, we are encouraged to use more simple words and sentences to describe the content of the text. It's fun and challenging! (H3_G6SB)

> To me, the cued language wrote on the blackboard is very helpful for me to memorise the whole text. Retell is actually belonging to recite. (L1_G6SA)

In addition, similar as the activity preferences of the Grade 4 low achievers, Grade 6 low achievers also like recite and read aloud, which they thought memorisation and mechanical drilling would be helpful for developing their listening and speaking skills. These activities were regarded relatively easy to participate in due to their limited English proficiency.

> I like reading aloud the text repetitively and volunteering to do it in class. From unfamiliar to familiar stage and then move on the reciting level, it could consolidate my understanding of the whole text and also help me with memorising more key vocabularies. (L2_G6SA)

When cross-checking the results of student questionnaire in Grade 6, the three assessment activities mentioned above in the focus group interview were all rated as most useful ones and thus consistent with the results in student questionnaire.

6.4 Summary of chapter six

This chapter reports research findings on student outcomes in Grade 4 and Grade 6. It aims to provide a snapshot of a sample of primary EFL learners learning outcomes at 2 key grade levels: 1) linguistic outcomes, or proficiency levels in oral production and aural comprehension in relation to the NECS and 2) non-linguistic outcomes or associated affective attitudes. In terms of linguistic outcomes, there are significant differences between Grade 4 and Grade 6 students' performance on two individual assessments. According to the results of panelist standard setting, there were about 38.68% Grade 4 and 45.63% Grade 6 students who achieved listening targets at Level 1 and Level 2 in the NECS, with 46.23% of Grade 4 students and 61.17% of Grade 6 students achieving the speaking targets at Level 1 and Level 2, respectively. As for the non-linguistic outcomes, the underlying factors in Grade 4 and Grade 6 students' affective attitudes towards EFL learning were identified via factorial analysis. Students' preferred assessment tasks were also found to vary across grade levels.

Chapter 7 Discussion and conclusions

7.1 Introduction

In this final chapter, the major findings are synthesised and interpreted. The significance of the study and its associated implications are discussed, limitations are indicated, and directions for future research are suggested.

7.2 Synthesis of the findings

The overarching aims of the study, as indicated in the introduction chapter, were to take a snapshot of how classroom-based language assessment was organised in the formal primary school setting in China. The study's primary aim was to examine the assessment of oral and aural language skills and their relation to the National English Curriculum Standards (NECS) targets. Specifically, the study comprised four main research questions, which, in turn, touched on three major objectives: (1) to go into the *black box* (i.e., the classroom) to investigate how teachers actually assess their students' oral and aural language skills according to NECS targets; (2) to explore the two stakeholders' (i.e., teachers' and students') attitudes to EFL teaching/learning and assessment; and (3) to examine students' actual achievements in listening and speaking via two external assessments and to interpret students' performance against the NECS targets by a group of local teachers through standard setting methods. In the following sections, the findings related to each research question are synthesised and discussed in turn.

7.2.1 The findings for RQ1: Inside the *Black Box*

This section summarises the research findings for the first main research question: *How do EFL teachers from Grade 4 and Grade 6 in these state primary schools conduct classroom-based listening and speaking assessments in relation to NECS targets?* The discussion aims to portray what was actually happening in the classroom in terms of assessing students' oral and aural language skills in relation to NECS targets. Specifically, the discussion aims to show which types of classroom-based assessment activities were being used by EFL teachers at these two grade levels (i.e., Grade 4 and Grade 6); and how these assessment activities were being implemented. The section synthesises the findings from both quantitative (i.e., the teacher questionnaire) and qualitative data (i.e., classroom observations and teacher interviews) and discusses the commonalities and differences that emerge across different strands of data.

7.2.1.1 Classroom-based oral and aural assessment activities in the EFL primary context

This section first summarises the commonly used classroom-based oral and aural assessment activities in the Chinese EFL primary context, specifically at two key grade levels (i.e. Grade 4 and Grade 6). The differences between these two grade levels are then identified and discussed.

Types of classroom-based oral and aural assessment activities

The most frequently used classroom-based assessment activities

Findings from the teacher questionnaire, which was distributed to a total of 103 EFL teachers from 28 local state primary schools, showed that the most frequently reported classroom-based assessment activities in use were (1) form-focused activities with a limited degree of communicativeness (e.g., mechanical read aloud; listen and imitate), assessing students' pronunciation and intonation; and (2) limited-production activities (e.g., Q&A led by teacher), which assessed students' understanding of simple questions or instructions and typically required short answers, either a word or a sentence in response.

Both types of assessment activities were also observed in the Grade 4 and Grade 6 classrooms of the two sampled state primary schools. Significantly more class time was devoted to these two types of assessment activities than to any other types of oral and aural assessment activity. Of the form-focused activities, read aloud at either the word or the phrase level was frequently observed in Grade 4 classrooms while recite text/dialogue was frequently observed in Grade 6 classrooms. Both kinds of activities were mechanical and sometimes decontextualised (e.g., read aloud) or memorisation-based (e.g., recite text/dialogue). Data from teacher interviews further confirmed that these traditional form-focused assessment activities were regularly integrated into daily classroom teaching at both grade levels. It seems that these activities were matched to those NECS targets focused on assessing students' pronunciation and intonation at the word and discourse level (i.e., to the first NECS target in speaking skills at both Level 1 and Level 2). While these activities assessed students' pronunciation in an isolated context, they failed to assess pronunciation in the context of language use, as per the central criterion of "intelligibility" suggested by McKay (2006, p. 190). That is, because these activities involved brief responses, they could not assess the intelligibility of students' pronunciation in discourse situations, in which words' pronunciations change in relation to their combinations into sentences and sentences' rhythms and intonational contours change in relation to their combination into clusters at the discourse level. Therefore, assessing students' pronunciation in an isolated context cannot identify the actual pronunciation errors that would be found in real-life interactions.

Regarding classroom-based assessment activities with limited-production, teachers from both grade levels were frequently observed using Q&A to assess students' oral interaction skills. Specifically, Q&A led by the teacher was frequently used by Grade 4 teachers whereas Q&A led by the group was mainly found in Grade 6 classrooms. This finding was also confirmed by the teacher interviews, which provided further details about the implementation of Q&A in the classroom. Namely, Grade 4 teachers reported that they mainly used restricted Q&A, along with occasional textbook role-play, to elicit one-word responses (i.e., vocabulary) and formulaic expressions from students. Grade 6 teachers, on the other hand, reported that they tended to use extended Q&A, which focused on information exchange tasks, and was led by either the teacher or the group. In both grades, the goal of the Q&A activity was to assess students' communicative language use in a real-classroom context. These activities integrated both listening and speaking skills and corresponded with the NECS targets emphasising communicative skills. However, the restricted Q&A in Grade 4, though in the form of a communicative assessment activity, was actually assessing students' memorisation of simple, learned Q&A exchanges that were routine in nature. Similarly, the textbook role-play also acted as a means of assessing students' rote learning of textbook dialogues, rather than the students' oral interaction abilities. Therefore, such assessment activities did not help either to establish whether students could use the target language communicatively or to demonstrate whether students could fully achieve the NECS targets related to communicative skills.

The least frequently used classroom-based assessment activities

The teacher questionnaire indicated that the least frequently used types of assessment activities were those assessing students' narrative skills (e.g., storytelling and picture description). Although the findings revealed that textbook retelling was reported to be frequently used (i.e., around once a week) by local EFL teachers, this activity belonged with the cued narrative tasks, in which students were asked to retell a text based on the visual aids and key vocabulary words or phrases provided. Such activities rarely gave students the opportunity to use their own target language to describe or to narrate creatively and spontaneously. In comparison, other narrative skills related activities like storytelling and picture description allowed students to use their own target language in a creative and spontaneous fashion. The findings from the classroom observations and teacher interviews also confirmed that textbook retelling was the only activity which both Grade 4 and Grade 6 teachers used for assessing students' narrative skills. No other forms of storytelling or picture description were observed. In addition, the class time allotted to textbook retelling was less than the time allotted to any other kind of assessment activity in Grade 4, meaning that narrative skills related activities received the least amount of coverage in Grade 4. While a relatively larger proportion of class time was devoted to textbook retelling in Grade 6, the overall percentage of class time was still lower than that allotted to communicative related assessment activities (i.e., Q&A) at both grade levels. Taken together, both the low frequency of use

and the comparatively limited class time devoted to narrative skills related assessment activities, the findings suggested that, narrative skills were not fully assessed during daily classroom teaching at either grade level. In addition, the types of assessment activities chosen also limited students' free use of their target language resources.

Features of classroom-based oral and aural assessment activities

The study also investigated the four features of classroom-based oral and aural assessment activities: these are participant organisation, intended content focus, student modality, and materials used. Below the similarities among the features found in all four EFL teachers' classrooms are discussed, as are the differences between the features used predominantly in assessment activities found across the two grade levels.

There were three salient features commonly found in the classroom-based assessment activities practiced by all four EFL teachers at two grade levels: these activities were organised as teacher-centred; they integrated listening and speaking skills and they were textbook-dominated. These findings revealed several important points. First, in terms of *participant organisation*, over 70% of the time devoted to assessment activities was organised by the teacher in terms of whole-class participation. In contrast, relatively little time was allotted to group activities, even though group activities were considered by the teachers to be essential for developing students' communicative skills, and the teachers acknowledged that such activities engaged students in genuine interactions (Spada & Fröhlich, 1995). In terms of *student modality*, for most of the assessment activities, both listening and speaking skills were assessed in combination. For example, the most frequently used activities during class, such as Q&A, used both listening and speaking skills to assess students' ability to interact. In contrast, activities that assessed only one set of skills, that is, either listening or speaking skills, were relatively fewer in number and frequency in the four teachers' classrooms. In terms of *materials used*, all four teachers used the textbook as a source, especially, for extended written text. In comparison, the teachers did not make as much use of activities involving minimal written text; when they did, these activities primarily involved isolated sentences or vocabulary words. Few authentic visual or audio materials were used either. Thus, three of the four features of assessment activities were similar across grade levels. Only one feature, *the intended content focus*, differed between the grade levels. Specifically, the assessment activities in Grade 4 were found to be largely focused on assessing the meaning expressed whereas Grade 6 activities targeted both language accuracy (i.e., linguistic form) and intended meaning; thus, in Grade 6 the activities were both form-focused and meaning-focused.

Discussion

This section summarises the findings regarding the types and features of classroom-based assessment activities by synthesising the data from the teacher questionnaire, classroom observations and teacher interviews.

The most commonly used oral and aural assessment activities at both grade levels were 1) the form-focused activities (e.g., read aloud; listen and imitate), which assess pronunciation and intonation; and 2) the activities involved with limited production (e.g., Q&A led by teacher or group), which assess communicative skills. In addition, cued narrative activities were frequently used in the upper grade level (i.e., Grade 6). Concerning these assessment activities' relationship to NECS targets, it was found that the activities frequently assessed students' listening comprehension and oral production skills at both the vocabulary level and the sentence level; however, students' skills at the discourse level were relatively less frequently assessed.

The use of controlled and guided assessment activities in the current study is consistent with the results of a large-scale survey study conducted by Brumen, Cagran, and Rixon (2009). That prior study (see previous discussion in section 2.4.2.2 in chapter 2) investigated the assessment practices of primary EFL teachers in three Eastern European countries. The results indicated that the most frequently used assessment activities were ones assessing interactive skills (e.g., oral dialogue; gap-fill activities) and oral fluency (e.g., repeating and drilling vocabulary words and sentences); in contrast, the assessment activities that required production at the discourse level and/or involved the creative use of language (e.g., learner-created sentences or texts) were seldom used. Such consistency in engaging in cognitively simple assessment activities in the classroom might help explain why students at the initial learning stage tend to memorise formulaic expressions and to produce language in whole chunks rather than in analysable components (Ellis, Brewster, & Girard, 2002; Skehan, 1998). As a further outcome, this state of affairs might help explain why some young learners, especially in an East Asian context, might be challenged by activities assessing students' creative use of language and their production or comprehension at the discourse level; the large class sizes and the mixed levels of learning ability found among the students might impede full implementation of such activities (Brumen et al., 2009; Wang, 2009)

As for the most common features of the classroom-based assessment activities, the current study revealed that these activities were organised by the teacher to the whole class and that the type of materials especially for extended written text, was typically the textbook. The current study also revealed that a traditional teacher-centred, textbook-based teaching approach was still common in the current assessment activities at the primary level in China. This conclusion is consistent with the findings of prior studies conducted at the primary level (Cheng, 2011; Goto Butler & Lee, 2010;

Prapaisit De Segovia & Hardison, 2008; Wang, 2009). However, the current study also takes note of some additional issues and changes, thus distinguishing its findings from those in prior wok.

First, whereas previous studies described the teaching context in generalised terms, this study examined the teaching context in detail. It took into consideration specific features of assessment activities, such as the organisation of participants. For example, the current study explored whether organising more group assessment activities (as opposed to teacher-centred ones) would be more beneficial for developing students' oral competence and for eliciting students' output, especially for beginning language learners at the lower grade level (e.g., Grade 4), whose output might be limited to learned patterns and routines. The study also considered whether teacher-centred assessment activities might be best served as scaffolding to assist students with understanding the main goals of a certain task, with the teacher then conducting observations and alternations during interactions with students. As was reflected by the teacher interview data, as well as the classroom observation data, classroom-based implementation of Q&A normally followed a sequence in which the displayed Q&A was initially modelled and led by the teacher prior to Q&A being led by a group. Furthermore, it was found that the proportion of group-led Q&A was greater in Grade 6 (Ms L: 7.5%; Ms B: 11.7%) than in Grade 4 (Ms P: 7.9%; Ms W: 2.9%). These findings might indicate that, as the students' proficiency increases, group-led Q&A is more likely to be effective in the upper grade level, especially when it is used to elicit students' genuine interactions and to focus on the meaning expressed instead of the linguistic form (i.e., the accuracy of the utterances) (McKay, 2006; Spada & Fröhlich, 1995). These findings were also consistent with the results of Butler and Zeng (2011)'s study, which indicated that the individual assessment format (i.e., student-teacher pair) could be beneficial for students with lower proficiency; however,. as borne out by the authors' later study, the group/paired assessment activity might have limited application for students from a lower grade level (i.e., Grade 4 in their study) (Butler & Zeng, 2014). Therefore, it might not be appropriate to polarise teacher-centred versus learner-centred approaches in relation to assessment activities. Rather, the two approaches might be seen in a continuum, considering the merits of both in relation to the students' level of proficiency as well as the assessment activity's objectives (Groundwater-Smith, Ewing, & Jean Le Cornu, 2006).

Second, unlike previous studies, whose analysis of assessment activities featured Chinese EFL classrooms dominated by activities like mechanical repetition and imitation of single words or sentences in an isolated context (Deng & Carless, 2009; Wang, 2009), the current study revealed that the majority of the Grade 4 teachers' assessment time was dedicated to meaning-focused assessment activities and that more than half of the assessment time in Grade 6 was devoted to form-meaning combined activities. In addition, the current study found that, at both grade levels, the main type of assessment materials used, though textbook-oriented, was extended, not minimal text. This finding might reveal that, at the activity level, students' aural and oral skills were assessed (1) with the intended focus on the expression

and negotiation of meaning, not on the accuracy of form and (2) mainly in a controlled and guided context, not in isolation. However, local EFL teachers' actual implementation of these activities are summarised and discussed in the next section.

7.2.1.2 Implementation of the classroom-based oral and aural assessment activities

This section summarises the findings regarding how local EFL teachers actually conducted the reported and observed assessment activities in the classroom. Specifically, two main agents of assessment are synthesised and discussed at two key grade levels (i.e., Grade 4 and Grade 6): teacher assessment and peer assessment.

Teacher assessment

The results of the teacher questionnaire indicated that activities involving the teacher as the main agent of assessment dominated in almost every lesson. This finding was later confirmed by the classroom observations, which presented direct evidence that over 80% of the assessment activities were evaluated by the teachers in Grade 4 and Grade 6. In addition, the classroom observation data further revealed that teachers primarily assessed individual student responses; whole-class choral answers and group responses were less frequently assessed. Nonetheless, as was elaborated upon in the teacher interview, giving opportunities to active, high-achieving students and nominating low-achieving students to participate in the assessment activities would help teachers better identify whether the students have achieved the learning targets or not.

Regarding the actual focus areas during the process of classroom assessment, the results of the teacher questionnaire revealed that while the teachers prioritized the students' correct use of pronunciation and intonation, the teachers also considered the students' appropriate use of vocabulary words, the students' affective attitudes (e.g., confidence and motivation), and the content or topic of the students' expressed responses. In contrast, this quantitative finding also revealed that the surveyed teachers reported that they treated the correct use of oral grammar as the least important focus area during the process of assessment. In comparison with the results from the classroom observations, the actual focus of assessment was reflected by the content of the feedback. Namely, both Grade 4 and Grade 6 teachers mainly provided form-meaning combined feedback. Their feedback focused not only on linguistic accuracy but also on the expression of meaning inherent in the students' responses. In Grade 4, around half of the assessment activities were provided with this kind of feedback; the proportion was even greater in Grade 6, where the form-meaning combined feedback was observed in the majority of the assessment activities (around 70%). Furthermore, the results revealed that this kind of feedback was usually provided in the combinations of the following forms: (1) in corrective form, by correcting linguistic issues such as pronunciation, vocabulary and grammar; (2) in confirmative form, by repeating

students' responses and/or (3) in explanatory form, by highlighting the information that students conveyed. It seemed that this form-meaning combined feedback at both grade levels balanced the focus on language use with a focus on linguistic forms. However, the analysis of the feedback interactions between teachers and students during the assessment process indicated that, even though Grade 4 teachers tended to use convergent follow-up questions for communicative purposes, the questions' format limited students' responses to simple yes or no questions. Similarly, Grade 6 teachers' feedback, although formulated in a combined form, placed greater emphasis on the accuracy of linguistic forms than on the expression of meaning.

Peer assessment

Peer assessment was the other reported and observed type of assessment in the current study. The results of the teacher questionnaire indicated that peer assessment was reported to occur once a week. This was also confirmed by later classroom observation: students' acting as the agent of assessment to evaluate their peers' performance constituted a small portion of the assessment activities. At both grade levels, peer assessment was usually teacher-led and commonly found in group activities like textbook role-play and Q&A led by the group. As with teachers acting as the agents of assessment, peers tended to give feedback to individual students. The content of this feedback was also largely limited to form-related comments, especially about correcting pronunciation, along with some holistic remarks (e.g., "Great!"; "Good job!").

Discussion

To sum up, this section depicted how local teachers actually implemented assessment activities in the classroom.

Overall, it was observed that the teacher acted as the main agent during the process of assessment, making judgements regarding students' performance. This finding was consistent with prior studies of young language learners in the East Asian context; those studies also found traditional teacher-centred learning, in which the teacher was seen as the sole authoritative figure and assessor (Wang, 2009).

Regarding the target of assessment, the current study revealed that teachers tended to focus more on individual students' responses than on choral answers involving the whole class or the group. This conclusion seems different from the findings of prior studies of China at the primary school setting; those studies found that teachers frequently used choral answers during teaching and assessment activities, a practice that might suggest the teachers' assumption that all students could learn at the same pace (Carless, 2001; Q. Wang, 2007). However, with the large class size and with mixed-level students, targeting choral answers could disguise individual students' learning difficulties and needs and would

not give the teacher an idea of the students' level of comprehension during the lesson (Holliday, 1994). In comparison, targeting individual responses could help teachers to identify individual students' weaknesses and provide individuals with more specific feedback (Mihaljević Djigunović, 2009). This approach to assessment would, in turn, facilitate students' learning and help the teachers to adjust their pedagogy accordingly. In addition, the current study also identified teachers' strategies for eliciting responses from students at different levels of proficiency. By engaging both high- and low-achieving students in assessment activities, local teachers could better assess whether students had achieved the learning objectives or not. However, as was stated by the local teachers, such strategies also carry negative consequences, since mid-level students might be overlooked. Again, this finding might be constrained by contextual factors, like the large class size and limited instructional time, which are discussed in more detail in section 5.2.2.2.

The current study also investigated the actual focus of the assessment as it was in progress, finding that most of the observed feedback was task-referenced, not ego-referenced. As discussed in the existing literature, task-referenced feedback is focused on students' actual performance in relation to specific activity requirements or the learning goals (Tunstall & Gipps, 1996). Thus, it is focused on the outcomes produced. In contrast, ego-referenced feedback is non-specific and non-evidence-based; usually, it takes the form of praise or judgement and makes comparisons between students and their peers. Thus, it is focused on the needs of the student and the student's affective attitudes, rather than on the outcomes of a task. Ego-referenced feedback could have a negative impact by redirecting a student's learning orientation towards ego-oriented rather than task-oriented outcomes (Black & Wiliam, 2009; Butler, 1988), and thus away from outcomes. Therefore, this finding indicated that local EFL teachers used students' learning evidence to promote students' immediate learning through task-referenced feedback. Furthermore, this task-referenced feedback could be further characterised as combined form-meaning focused, meaning that it seems balanced focus on language use and linguistic forms.

However, the results revealed that the confirmative form of the feedback and the convergent nature of the follow-up questions were usually found in Grade 4 teachers' feedback, whereas an emphasis on the accuracy of linguistic forms was more pronounced than attention to language use or meaning in the Grade 6 teachers' feedback. This finding indicated that the teachers' actual focus (as opposed to their perceived or reported focus) might direct students' engagement in the assessment activity. For example, the actual focus of the textbook retelling in Grade 6 was on memorising the text accurately, not on developing students' narrative skills to express meaning. Grade 6 students might, therefore, treat this activity as another way to check their memorisation of the text. This finding confirmed earlier findings in Butler (2005)'s study that even when teachers employed the same communicative activities, the actual activities that students engaged in appeared to be very different depending on the actual enactment by the teacher (Butler, 2005).

In addition, the current study also identified that peer assessment was implemented and embedded within the teachers' assessment, especially during group activities. The nature of this peer feedback tended to be voluntary, centred on holistic comments, and limited to the correction of pronunciation. These characterisations might be due to the students' limited language proficiency, especially in the current EFL primary context. However, the potential benefits of such peer assessment were also noted in the current study. Students were encouraged to actively engage in the assessment process to evaluate their peers' performance. Furthermore, the very act of providing feedback could be seen as another opportunity for students' language use as well as a supplement for teacher assessment at a lower cost (McKay, 2006; Patri, 2002). However, the content of peer feedback might also reflect teachers' actual orientation during the assessment process: as the data from the teacher questionnaire and the classroom observation also bore out, the correct use of pronunciation and intonation was, in fact, the primary focus during oral and aural assessment activities; not surprisingly, then, the other aural and oral targets in the NECS seemed less frequently covered during peer assessment also.

7.2.2 The findings for RQ2: Factors related to classroom-based assessment implementation

This section summarises the research findings for the second main research question: *How is the intended national curriculum being interpreted by local EFL teachers?* This question aims to explore any factors that, from the teachers' perspectives, either facilitate or inhibit implementation of the NECS in the classroom. 103 EFL teachers from 28 state primary schools in the local district were asked to respond to the teacher questionnaire; in addition, four EFL teachers from Grade 4 and Grade 6 were interviewed separately after classroom observations were conducted. The analysis of the survey data resulted in the discovery of four significant factors: (1) external support; (2) contextual constraints; (3) testing washback and (4) teachers' oral proficiency. Meanwhile, the analysis of the teacher interview data identified four emergent themes: (1) knowledge and understanding of the NECS; (2) contextual conditions; (3) teacher factors; and (4) resource support. A certain degree of both overlap and divergence was found regarding the teachers' general perceptions toward NECS implementation. This section synthesises the findings from these two strands of data and discusses the two main factors (i.e., external and internal factors) that emerged across the data strands.

7.2.2.1 External factors

<u>Available resources and support</u>

External support was the first factor obtained from the factorial analysis of the teacher questionnaire data. It incorporated two initially designed factors, which were teaching resources and teachers' professional development. With the rounded mean of 4 on a scale of 5, this factor indicated that teacher respondents agreed that the current external support, including the NECS and the textbook, as well as

the teacher training, facilitated their classroom EFL teaching. This was also confirmed by the teacher interviews, in which local teachers also recognised the fundamental role of the NECS as a rough teaching guide and acknowledged the benefits of both in-school and local government trainings. However, the interviewed teachers also felt that the content of the NECS was ambiguous; the difficulty level of the textbook was inconsistent across the grade levels; and the training support at the national level was impractical.

Fairly recent studies reported similar findings regarding external support. As was suggested by Cumming (2009), curriculum standards were established based on professional judgment or consensus, not empirical data or theoretical frameworks. The implication is that empirical data or theoretical frameworks are needed to make evidence- or theory-based decisions. The lack of such a process has contributed to the relatively imprecise wording of standards, which, in turn, leads to challenges for classroom teachers relying on these standards both as a de facto teaching guide and as a set of objectives or goals for their students. The impracticality of teacher trainings at the national level also likely reflected that, at this level, a top-down approach was adopted to the interpretation of the curriculum and that this approach did not often meet—or perhaps consider—local needs (Butler, 2015; Zhang & Adamson, 2007). Textbooks also posed additional difficulties. As the core resource that most directly reflected the principles of the curriculum standards, the textbook was often viewed as Bible by local teachers, who believed that everything that appeared in the textbook had to be taught and learned (Butler, 2015; Wang, 2009). However, as was revealed in the current study and was also previously noted in Li and Baldauf (2011), inconsistencies in the difficulty levels of the chosen textbooks from one grade to another likely caused confusion and pressure for local teachers striving to cover everything in the textbook, let alone to include other authentic materials or language-use activities during class.

Contextual conditions

With regard to contextual teaching conditions, the questionnaire data also confirmed that the *large class size* and *insufficient instructional time* were reported to be major constraints impeding teachers' curriculum implementation, especially for teaching and assessing oral-related skills. These factors had been identified in many prior studies as common characteristics of formal instructional EFL learning contexts in East Asian countries (Carless, 2003; Enever, 2018; Nunan, 2003; Su, 2006; Wang, 2009). With around 50 to 70 students per class, 3 to 4 classes to teach, and only 2 to 3 class hours available each week, overloaded EFL teachers were challenged to design and to conduct oral related activities to assess students' oral language proficiency. These same EFL teachers were further challenged to design and to conduct oral related activities in such a manner that all students, irrespective of their ability levels, could fully participate in these activities and would be fully motivated to do so (Murphy, 2014). Not surprisingly, the current study found that learners' opportunities for language contact were fairly limited

at both grade levels, even though a number of group assessment activities were observed, especially in Grade 6 (Enever, 2018).

Regarding the *washback effects of exams*, the questionnaire data, with a mean score of 2.83, confirmed the findings from the interview data that local teachers' instruction of listening and speaking skills tended to be less influenced by exams. This finding seems different from prior studies; in those studies, teachers in East Asian countries were commonly believed to experience significant testing pressure, and this pressure was believed to exert considerable influence upon their teaching and assessment practices (Butler, 2015; Deng & Carless, 2010; Wang, 2009). In the current study, it was found that the local government set the end-of-term exams as low-stakes, likely causing the positive washback effect on teaching and learning. Classroom observations confirmed this impression since most of the oral and aural assessment activities were oriented for language use; they did not take the form of traditional paper-and-pencil tests. However, in teacher interviews, an exception to this generalisation was found for teachers at the upper grade level (i.e., Grade 6), who reported that their teaching practices were still being influenced by junior entrance exams. Students responding to parental pressure were more willing to focus on test-related content and on test-taking skills in order to achieve higher scores on entrance exams. This finding revealed that, even though examinations at the primary levels seemed to be low-stakes, the examination-driven educational system, overall, remains unchanged in China (Deng & Carless, 2010; Li & Baldauf, 2011).

7.2.2.2 Internal factors

Teachers' oral proficiency, as an internal factor, was highlighted in both strands of data. In the teacher questionnaire, respondents showed a certain amount of confidence in their English language proficiency, with a rounded mean of 4 on a scale of 5. It might be important to note that the questionnaire items (e.g., I can use English to discuss with students) under the factor scale of *teacher proficiency* focused on teachers' perceived language use ability in classroom-based contexts and with classroom-based tasks, as opposed to their language use ability more generally (i.e., in real-life contexts or freeform oral interactions). Classroom observations further revealed that teachers relied heavily on the textbook, using fixed target sentences and vocabularies from the textbook to interact with students in predictable formats, for example, during Q&A sessions. The teachers' feedback to students also was more concerned with the correction of students' pronunciation and vocabulary. Follow-up questions were also framed in convergent forms featured by closed questions with the intention of eliciting expected answers (Torrance & Pryor, 2001). These features might explain why the teachers perceived their language proficiency to be at a relatively satisfying level since the communication between teachers and students went smoothly, with each participant taking turns. However, the nature of the interactions actually "limited the amount and quality of information afforded for assessment purposes" (Hill &

McNamara, 2011, p. 10). In fact, during the teacher interviews, local teachers acknowledged their lack of sufficient English proficiency to conduct authentic communicative activities and to create a better language learning environment. Indeed, teachers' English language proficiency has been considered essential for effective English teaching in the EFL context. The findings of the current study were also similar to the results of a previous empirical study of EFL primary teachers in East Asian countries (Butler, 2004). That study indicated that EFL teachers' current proficiency levels fell below the minimum proficiency levels required, especially in the domains of oral production and oral grammar. Thus, teachers' English oral proficiency seems crucial to successful EFL instruction at the primary level for several reasons. Not only do most classroom activities at this level involve oral interaction, but curriculum standards also emphasise oracy skills, which were also first developed for young EFL learners.

In addition, the results of the teacher interviews indicated that, given these external factors, the degree to which teachers minimized the influence of contextual constraints and made full use of the limited resources depended on the teachers' *professional pedagogical knowledge and skills*. This seems still to be a worldwide issue today: there is still a shortage of qualified EFL teachers at the primary level and a lack of the professional trainings and systematic supports needed to equip these teachers to improve both their English proficiency and their practical knowledge (Brumen et al., 2009; Butler, 2015; Chou, 2008).

7.2.3 The findings for RQ3 and RQ4: students learning outcomes by Grade 4 and Grade 6

This section summarises the research findings for the third and fourth main research questions: *RQ3. Upon finishing Grade 4 and Grade 6, respectively, what can young EFL learners actually do/achieve in speaking and listening and how do these skills accord to the targets set by NECS*?; and *RQ4. What are the attitudes of young EFL learners towards learning, teaching and assessment of EFL in Chinese primary state schools?* Both of the research questions are focused on Grade 4 and Grade 6 students' learning outcomes; the first question focuses on linguistic outcomes, and the second focuses on students' perceptions of their learning experience. The aim of these questions was to examine (1) what students can actually do in terms of listening and speaking via two external assessments; (2) how a group of local teachers interpreted their students' performance in relation to the NECS targets; and (3) what affective attitudes students held towards EFL learning and how the students viewed classroom-based aural and oral assessment activities. For the linguistic outcomes, 106 Grade 4 students and 103 Grade 6 students were individually assessed, and their results were linked with the NECS targets via the interpretation of a panel of local expert teachers. The non-linguistic outcomes were obtained from the student questionnaire and focus group interviews. This section summarises the findings of the students' learning outcomes and discusses the related issues that emerged from the study.

7.2.3.1 Students' linguistic outcomes

Descriptive results

In this study, students' aural and oral linguistic outcomes were measured via two externally developed assessment instruments: the British Ability Scale-II (BAS-II; Elliot, Smith, & McCulloch, 1996) verbal comprehension test and a storytelling task based on the book *Frog, where are you*? (Mayer, 1969). In the absence of any standardised assessment instrument based on the NECS available for use at the primary EFL level in China, the constructs of these two assessment instruments (see section 3.4.4.1) were assumed to be matched with the level descriptors in the NECS and to be appropriate for use with both NECS levels (i.e., Level 1 and Level 2). The results of the BAS verbal comprehension revealed that Grade 6 students' listening comprehension was significantly higher than their Grade 4 counterparts'. The results of the storytelling task indicated the following: 1) at the macrostructure level, the essential story grammar elements (i.e., initiation event, character, action and consequence) were found in both Grade 4 and Grade 6 students' narratives. In addition, more complex details emerged in the Grade 6 students' narratives, and these elements showed a significant growth in the Grade 6 sample; 2) at the microstructure level, Grade 6 students' oral narratives exhibited a longer output with a richer vocabulary and with more complex syntactic structures than was found in the Grade 4 students' oral narratives.

Linking results

In order to interpret and map the descriptive results of the two assessment instruments with the targets specified on the NECS, two standard-setting methods were applied in the current study to construct the cut-off scores for the two assessment instruments corresponding to the two levels of the NECS. Specifically, the modified Angoff Method was used for the BAS-II verbal comprehension and the contrasting group method was used for the frog storytelling task. The cut-off scores were constructed following after the two stages during the standard-setting procedures: familiarisation and standard setting.

During the familiarisation stage, local expert teachers (i.e., the panellists) developed a set of agreed-upon descriptors for listening and speaking skills to describe a just-qualified candidate at both Level 1 and Level 2 of the NECS. The results of familiarisation stage reflected that the panellists adjusted the original descriptors and added more details based on their familiarity with the students and their understanding of the local context. In comparison with the general descriptors of the NECS, some common characteristics were found among the panellists' descriptors across the two levels. First, the textbook was frequently added as a condition in the panellists' descriptors. For example, the panellists agreed that a minimally qualified candidate at Level 1 could recognise pictures or real objects based on verbal instructions using vocabulary words or phrases learned from the textbook. Such textbook-

oriented judgements (i.e., the treatment of the textbook as a Bible for both teachers and students) were also reflected by the current study's different data strands. All the assessment activities were closely related to the textbook content. Students thought highly of the textbook and showed a willingness to participate in textbook-related activities; low achievers, in particular, seemed to feel that mastery of the content of the textbook would help them to pass the exams. From the perspective of these local expert teachers, then, the textbook seemed to hold primary importance in China and to serve as the anchor or baseline for interpretation of the NECS targets (Wang, 2009).

Second, the teacher was also frequently mentioned in the panellists' descriptors, even though the general NECS target statements do not reference the specific interlocutor in speaking or listening tasks. It could be the teacher or the peers. However, the panellists took teacher as the default and the only one when developing the descriptors. For example, the panellists agreed that a just-qualified candidate could understand and respond appropriately to familiar instructions and questions from their own classroom teachers. The perceived need for this descriptor could reflect that the teacher still dominated the class, a finding that was confirmed by the teacher questionnaire as well as by the classroom observations. Importantly, what is left out of consideration is how a just-qualified candidate could interact with peers.

Turning to the standard-setting results, based on the cut-off scores established for the two tests, fewer than half of the students in Grade 4 (38.83%) and Grade 6 (45.63%) were judged to have achieved the NECS listening targets at Level 1 and Level 2, respectively. With regard to speaking skills, fewer than half of the Grade 4 students (46.23%) were classified as qualified candidates at Level 1, but more than half of the Grade 6 students (61.17%) were judged as qualified candidates at Level 2. These quantitative judgements contrast with the qualitative predictions made by EFL teachers in the current study. In this study, all EFL teachers in the local district were asked to estimate the proportion of their students that was likely by Grade 4 and Grade 6 to reach the NECS targets in listening and speaking skills. The results of the teacher questionnaire revealed that: (1) for the listening targets, the majority of the students (75%) were estimated to achieve most of the listening targets at Level 1 and Level 2, while only around half of the students were estimated to achieve the target for text-level listening comprehension; (2) estimations for the speaking targets at both levels were similar: 75% of the students were estimated to achieve the targets related to pronunciation and communicative skills at both levels, while around 50% students were estimated to achieve the target for narrative skills.

In comparison, there seemed to be a discrepancy between the teachers' estimations and the panellists' judgements. However, this discrepancy might be explained by the following reasons. First, the two assessment instruments were externally designed and were conducted by the researcher instead of the local teachers. As a result, the content and materials might appear unfamiliar to the students and that unfamiliarity could have affected the students' performance, even though during adaptation and the

pilot stage, the content of the assessment instruments was reviewed to ensure it matched with the NECS curriculum and the students' textbook. The two externally designed assessment instruments also did not cover all the listening and speaking targets in the NECS, and the oral assessment task (i.e., the frog storytelling task) primarily assessed students' narrative skills, which local teachers perceived to be relatively difficult for their students. Therefore, the percentage from the linking results might be lower than the teachers' estimations. Second, the teachers' estimations might be influenced by the support that is intrinsically provided to students within classroom-based assessment activities, even if the teachers themselves do not perceive these supports. As has been suggested by Liddicoat (1997), assessment tasks in a classroom context might be over-supported, and teachers' judgements regarding student achievement could be affected by a series of support-originated constraints. These constraints included the *teachers* themselves (although the fact that the teachers were providing constant support during the task might make it difficult for them to evaluate students' actual proficiency in an objective manner); the *routine situation*, in which language use was formulaic and predictable; and the *prior work*, in which language had been prepared or rehearsed beforehand and could not represent students' spontaneous language use (Liddicoat, 1997, P25). In the current study, most of the assessment activities were textbook-based and students' output mostly involved language chunks and formulaic expressions. However, these activities, in comparison with the externally designed assessment tasks, might not reflect students' actual language proficiency. Students' conventional activities and interactions with teachers in the classroom might give teachers the impression that the majority of the students could achieve the curriculum goals, when, in fact, the objective evidence presented by externally designed assessments suggested a lower proportion as actually doing so.

7.2.3.2 Students' affective outcomes

Students' general affective attitudes towards EFL learning

In the current study, findings regarding students' general affective attitudes towards EFL learning were obtained by means of a student questionnaire. The exploratory factorial analysis of the survey data aimed to identify the main constructs of Chinese primary EFL students' affective attitudes toward learning English and resulted in the identification of four significant factors, which were *instrumental motivation*, *intrinsic motivation*, *anxiety* and *confidence*. In addition, further comparisons were made between Grade 4 and Grade 6 students to see whether there were any changes in affective attitudes among these identified constructs over time. The results revealed that students from both grade levels displayed similar affective attitudes under these constructs. Students held relatively strong instrumental and intrinsic motivations toward EFL learning. Although the instrumental motivation was found to decline significantly in Grade 6, the rounded mean of 4 on the scale of 5 indicated that Grade 6 students still held a positive instrumental motivation for learning EFL. Declining motivation for EFL learning

had been identified as a common issue in prior studies. In these studies, as in the current one, findings reported students' initial motivation for EFL learning to be relatively high; however , due to contextual constraints (e.g., large class size, limited instructional time, unqualified teachers and lack of language contact outside school), some students developed negative attitudes toward EFL learning (Carreira, 2006; Lopriore & Mihaljević Djigunović, 2009; Muñoz, 2017; Nikolov & Mihaljević Djigunović, 2011; Szpotowicz et al., 2009). Regarding the two other significant factors, confidence and anxiety, both Grade 4 and Grade 6 students were found to hold neutral orientations; also, there was no significant difference found for these two factors between the grade levels. The findings were consistent with previous studies conducted at the primary level in East Asian countries, which had indicated that the level of anxiety and confidence did not vary much across the grade levels (Butler & Lee, 2006; Carreira, 2006). However, it should be noted that neutral orientation in anxiety does not indicate that students, including the young learners who were commonly believed to treat EFL learning as anxiety-free, enjoyable and fun, had no anxiety at all about EFL learning (Nikolov & Mihaljević Djigunović, 2011). As the reversed mean score of the anxiety factor was lower than 3, it indicated that there was still a slightly elevated level of anxiety found among Chinese EFL young learners at both the lower and higher-grade levels.

Students' attitudes and views towards the aural and oral classroom-based assessment activities

Students' attitudes and views, specifically towards their EFL oral and aural learning, were explored in-depth via a smiley face questionnaire and focus group interviews. Students' voices are important since they are the primary stakeholder who experiences classroom activities. Thus, the focus group interview format was significant for the current study because it enabled students to reveal their thoughts and feelings about their learning experience in the classroom (Enever, 2011); however, it should also be recognised that, due to their young ages, the children could convey views and opinions that were unstable (Enever, 2011). Therefore, the findings from the group interviews and the smiley face questionnaire were discussed in reference to the multiple data strands obtained from teachers via the questionnaire, interviews and classroom observations.

The results of the student questionnaire indicated that, in general, at both grade levels, students held positive attitudes towards all the classroom-based aural and oral assessment activities. However, declining affective attitudes were detected between grade levels. Of the 18 total assessment activities, Grade 4 students held a strong orientation towards the majority of the activities, whereas Grade 6 students' preference declined for half of the activities. Among the students' top most preferred activities, the only activity that overlapped between Grade 4 and Grade 6 was listening comprehension in the traditional paper-and-pencil format. This finding might be interpreted as evidence that the washback effect of traditional examinations still existed; it could also be interpreted to demonstrate that this format

of activity for assessment of these skills existed even though it was not encouraged in the NECS. Students' focus group interviews also revealed that for both high and low achievers, they showed strong motivation to participate in such listening comprehension exercise in preparation for the future potential high-stake exams. In addition to this commonality, student's preference towards the assessment activities at two grade levels displayed different pictures. For the Grade 4 students, the students' most preferred activities were those which could get students actively participated in and involved with more chunks of language and formulaic expressions directly from the textbook, for example, textbook role play and textbook retelling. This is in line with Skehan (1998)'s view that young learners tended to apply their formulaic system in which they relied heavily on accumulating chunks and draw on the resources of the formulaic items to join the communication to get their meaning across. This was even more frequently mentioned by the group of low achievers. The focus interview data further revealed that low achievers show particular interest in the form-focused assessment activities like dictation, recite and read aloud. Though treated as boring ones for the high achievers, the low achievers' preference might reflect their weak foundations in the basics of the language (e.g., pronunciation and vocabularies). Therefore, it would be suggested that, in the EFL context, especially in the large class size with varied language abilities, teachers should be aware of students' different needs and provide equal opportunities for both low and high achievers to participate during the assessment activities.

For the Grade 6 students, the most preferred activities were those involved with genuine communications and information exchange, for example, Q&A led by the teacher and make up conversations. This reflected that as with student's growing proficiency, Grade 6 students become consciously aware of the underlying rules of the language in order to precisely express their meanings or pass information (McKay, 2006; Skehan, 1998). The data from the focus group interview also confirmed their preference and motivation in participating in such assessment activities. However, the positive perceptions of high and low achievers displayed different angles. High achievers enjoyed the challenge of creatively using the target language and interacting with their teachers. While for the low achievers, they treated it as a good opportunity to get feedback, especially on form-focused aspects (e.g., correcting pronunciation), from their teachers and to consolidate the pre-learned words and phrases by reciting the sentences from the textbook. It seems that the low achievers still heavily relied on their formulaic system to learn the language through memorisation and their rule-based system has not been effectively activated (Skehan, 1998). Moreover, it could be noted that the proficiency distance between low and high achievers was becoming wider in the upper grade level.

Furthermore, it could also be noted that the student's most preferred assessment activities were also overlapped with the most frequently used assessment activities observed during the class. This might suggest that teacher's practice and belief would also exert influence on student's belief (Zhang & Head,

2009). Moreover, teacher's choice of assessment activities and teacher's actual focus during the assessment would also affect student's perceptions towards EFL learning.

7.3 The significance and implications of the study

Based on the findings discussed above, this section draws on the significance and implications of the current study. In brief, the current study's significance and its associated implications are threefold: there are implications for pedagogy; implications for policy and curriculum and implications for resources and teachers' professional development.

7.3.1 Implications for pedagogy

Given that only a few prior studies had been conducted in formal instructional EFL setting at the primary level in East Asian contexts, especially in China, the current study provided a local picture of what was actually happening inside the *black box* (i.e., the classroom). The study did so by providing multiple strands of evidence from the two primarily involved stakeholders (i.e., teachers and students) and by documenting frequently used types of aural and oral assessment activities at two key grade levels. These activities were identified by the core three-step criteria and analysed through the adapted COLT scheme. In terms of activities, the current study not only analysed their associated features, but it also examined the relation of these assessment activities to the NECS targets set at two levels (i.e., Level 1 and Level 2). In terms of practice, the current study also analysed teachers' actual implementation of assessment activities through the lens of teachers' feedback to students to evaluate the extent to which the intended focus of assessment matched with the students' actual learning outcomes. This research design provided a dynamic angle to evaluate assessment practice. Therefore, the pedagogical implications were provided at two levels: both activities and practice.

7.3.1.1 Pedagogical implications at the level of activities

In terms of aural and oral classroom-based assessment activities, the results of the current study indicated that students' proficiency was mainly assessed through activities that unfolded in a controlled and guided way and that involved limited production on the part of both the teacher and the students (e.g., restricted Q&A). As a result, these activities helped to assess only whether students had memorised the language taught; the activities presented fewer opportunities to assess whether the students could use the target language creatively or not. Thus, even while it can be acknowledged that, at the initial stage, young learners at the lower grade level relied on a formulaic system to employ pre-learned language chunks and simple learned Q&A routines to answer questions or narrate (Skehan, 1998), EFL teachers should consider either creating new assessment activities or, at least, modifying existing ones in ways that will activate students' formulaic and rule-based systems and thus engage

these same students in oral interactions that go beyond the predictable and the routine. Because the NECS target focuses on language use in real-life contexts and that assessments delivering, if not on the face of it targeting, routine exchanges are not, in fact, well designed to meet those targets. Since the most frequently used aural and oral assessment activities in the current context were in a Q&A format, teachers could transform the restricted format into an extended one, as had been previously suggested by McKay (2006), by considering the following five tips, which are to be implemented in alignment with students' increasing proficiency:

> (1) adding an element of surprise and unpredictability; (2) increasing the complexity of the questions (even through the expected response may remain simple); (3) including new vocabulary (sometimes unknown) into the questions; (4) supporting new language in questions with gestures, objects, and pictures to help children to predict from the context; (5) supporting new language by asking simpler or explanatory follow-up questions where needed. (McKay, 2006, p. 200)

It should also be noted that, although the above suggestions for extending the Q&A could help students to call on their language resources and help students use language spontaneously and creatively, students' affective attitudes, as well as their level of proficiency, must be taken into account. As was revealed by the results of the student questionnaire, although the students displayed positive intrinsic and positive instrumental motivations toward EFL learning, the students in both grade levels also displayed a relatively low level of anxiety. Therefore, in order to engage students in language use, the assessment activities also needed to provide both visual supports (e.g., puppets or pictures) and conversational supports (e.g., gestures and facial expressions) to maintain student interest, to alleviate anxiety and to help the students maintain concentration (McKay, 2006). Moreover, the results of the focus group interviews further indicated that in a large class size with mixed levels of language ability, high- and low-achieving students displayed different affective attitudes towards different classroom-based assessment activities. For example, in the current study, low-achieving students at both grade levels showed a preference for participating in assessment activities that involved routine exchanges and/or expected responses that could be found in the textbook (e.g., read aloud and recite text/dialogue), while the high-achieving students preferred assessment activities that involved unscripted interactions. Therefore, the pedagogical implication is that teachers should cater to the different needs of students and provide balanced opportunities to ensure that all students can experience a sense of accomplishment during the activities. The form-focused assessment activities thus should not be totally avoided since low-achieving students clearly still relied on their formulaic system to participate in activities, and these students also deserve opportunities to participate and practice.

7.3.1.2 Pedagogical implications at the level of practice

Second, regarding the level of practice, the analysis of the feedback reflected that, in the current study, teachers' actual focus and students' learning outcomes were not always matched with intended focus

of the assessment; for example, in Grade 6, teachers typically used linguistic accuracy as the criterion to evaluate students' responses and to judge the successful completion of assessment activities, even when these activities were intended to be meaning-focused or to have a combined focus upon both form and meaning. This finding revealed inconsistencies in teachers' approach to designing, operating and administrating the assessment activities. To help ensure a better alignment of the focus of assessment and students' learning outcomes, while also matching both better to the NECS targets, the current study thus recommends introducing the following set of questions found in Table 7.1. These questions address all three phases of assessment (i.e., design, operations, and administration) and can be used as a guide to follow throughout the classroom-based assessment process. During the design phase, to make sure that the assessment will be useful, teachers are encouraged to check the following: the curriculum targets, the characteristics of all learners in their classroom and the construct to be assessed (Bachman & Palmer, 1996; McKay, 2006). With these questions in mind, teachers could evaluate at the design phase whether the assessment would "give assessors the best evidence for assessment decisions about children's language use ability" (McKay, 2006, p. 113). During the operational and administration phases, teachers are encouraged to select the appropriate tasks and to make any necessary adaptions to ensure that the students' language abilities and needs are successfully checked and that the students' learning and motivation are both activated by the assessment activity. Moreover, teachers are also encouraged to ensure that students also understand the assessment objectives (i.e., what is required from their performance) and the evaluation criteria (i.e., what is expected of their performance). If not, the students might misunderstand the focus of assessment and thus aim for the wrong targets in the performance.

Table 7. 1 A set of questions for classroom teachers to follow during the process of assessment (McKay, 2006, p.147)

Design phase
Why do I need to know, and who else needs to know? (i.e., the purpose the assessment will serve)
What do I need to know? (i.e., the constructs that will be assessed)
Operational phase
How can I find out? (i.e., the tasks and strategies that will be used)
What will I do with the information? (i.e., how information will be used, assembled and stored)
Administration phase
How will I know that the assessment has been effective and how can I improve upon it next time? (i.e., the evaluation of the assessment process)

7.3.2 Implications for resources and for teachers' professional training

The in-depth empirical data obtained from the teacher questionnaire and from the in-depth interviews with four teachers further explored both the external and internal factors that might facilitate or inhibit local teachers' implementation of the NECS, especially with regards to the NECS aural and oral components. Identifying these external and internal factors might help all stakeholders in EFL

instruction to better understand the difficulties and challenges that teachers face during curriculum implementation in a local district in a city in China and thus to consider changes that could facilitate future curriculum implementation and thus potentially improve students' achievement of the NECS standards. The implications that follow from these findings are provided at two levels: external and internal factors.

7.3.2.1 Pedagogical implications for resources and support

With regard to external factors, the availability and quality of supplementary resources and support were highlighted by the current study. Although teachers recognised the local resources and professional trainings provided by the schools and local government, the results indicated that there was still a perceived gap and disconnection between the imposition of the NECS by the national government (which teachers perceived as a top-down initiative) and the interpretation of these standards at a local level in accordance with local needs (which teachers perceived as a bottom-up effort). On the one hand, there was still a reported lack of either detailed supporting documents or clear guidance from the national government to help the local government and its schools to further interpret the NECS and to effectively implement curriculum in accordance with the NECS targets; the local educational bureau and school teachers were left on their own to interpret curriculum standards and NECS targets as best as possible. However, such local interpretations might either fail to reflect the principles intended by the NECS or prove to be inconsistent with the standard terminology and conception of assessment defined in the existing literature. On the other hand, the available online national in-service trainings provided to local teachers were described by local teachers as impractical and ineffective; these trainings failed to equip teachers with the skills needed to further adapt and reinterpret the NECS according to local contexts. Therefore, an important pedagogical implication of the current study is that there is an urgent need for high-quality pre- and in-service training courses, together with supplementary materials that provide clearer definitions of NECS targets as well as detailed instructions on how to understand these targets and how to implement and assess them in daily classroom teaching. Moreover, as suggested by Kumaravadivelu (2001), the trainings and materials provided should be context-sensitive, meaning that they should follow the principles of "particularity, practicality and possibility", rather than directly employing Western approaches and materials, as is often the case in East Asian contexts today (Garton et al., 2011). According to this recommendation, first, the trainings should target the following objectives to improve teachers' assessment literacy: to enhance teachers' knowledge of and their skills in implementing various assessment activities and techniques; to develop teachers' ability to adapt and to use these assessment activities and techniques more efficiently (i.e., in ways that take into account contextual factors, such as large class size, mixed levels of language ability in the classroom and limited instructional time). Second, the teaching and assessment materials should be made available in various formats (e.g., paper, radio- or video-based) to accommodate varying local

conditions. The teaching and assessment materials should also be selected with the age of the learners and their affective needs in mind. In other words, the teaching and materials must take into account not only the curriculum targets (which is also crucial) but also the need to motivate students to learn English. Finally, in whichever formats these teaching and assessment materials are provided, there should be accompanying instructions to help local teachers use and adapt these materials effectively.

7.3.2.1 Pedagogical implications for the teacher professional development

With regard to the internal factors that might facilitate or inhibit local teachers' implementation of the NECS, teachers' proficiency was identified in the current study as a critical component for effective curriculum implementation, especially in the oral and aural domains. Here, the teachers' proficiency was not only related to their English language use ability, but it also referenced their professional pedagogical knowledge of the teaching and assessment of young learners. Both aspects—language proficiency and pedagogical knowledge--are indispensable. To improve teachers' English proficiency, the national and local governments should consider providing EFL teachers with long-term English language development training courses. These courses should also be accompanied by a systematic and ongoing evaluation system, enabling teachers to periodically assess their proficiency and appreciate any progress; the current system, i.e., a one-time evaluation of EFL instructors, does not encourage or provide a means to acknowledge growth in proficiency (Butler, 2004). Although the current study also discovered a trend of EFL teachers' holding a higher level of credentials (i.e., more than half of the surveyed teachers held a bachelor's degree in English), this trend does not guarantee that these teachers could necessarily retain their English proficiency over the span of their teaching careers if no further assessment were required.

In addition to these recommendations regarding improvements in English proficiency, EFL teachers also need to develop their professional pedagogical knowledge and skills. Although the pre- and in-service trainings and materials suggest the introduction of different types of assessment activities in the classroom, according to the Activity Theory (Lantolf, 2000), even the teachers used the same assessment activity and students appeared displayed the same behaviours, students might in fact be engaged in the different activities as teachers might establish different motives and goals for the same activity. Doing so might further result in different learning outcomes. Therefore, as suggested by McKay (2006), an effective method for improving teachers' classroom-based assessment skills is for teachers to work collaboratively together. Teachers working collaboratively as a group can help each other to identify goals that are both possible and feasible for the suggested assessment activities while also taking into account constraints posed by the local syllabus, the NECS targets and the contexts and needs of the local students. Then, based on this shared input and their collaboratively made decisions, the teachers could share ideas, plan procedures and try out adapted assessment activities together.

7.3.3 Implications for assessors

Thus far, the pedagogical implications discussed address the current study's focus on the assessment process that occurs in the classroom; however, the current study also sought to assess students' learning outcomes in relation to NECS targets. Indeed, this study represents the first attempt to use two externally designed individual assessment instruments to assess the learning outcomes of Chinese EFL primary students. The descriptive results indicated that this approach could make significant distinctions between the listening and speaking skills achieved by students in the lower and upper grade levels (i.e., Grade 4 and Grade 6). For the listening test, the BAS-II Verbal Comprehension (Elliot et al., 1996) was adapted and used. The selection of the test for the current study was based on its advantages for assessing students' listening skills. First, the test content was age-appropriate. Second, its procedure was child-friendly and familiar to young learners, who were used to performing similar tasks (i.e., either manipulating toys or pointing to pictures based on aural instructions) during daily classroom learning. The implication of this finding for future assessment is that there may be a benefit to assessment tasks that require a minimal oral or verbal response to demonstrate young EFL learners' understanding. Young learners enjoy responding with actions more than they do with language, and these actions may be elicited as evidence of their comprehension (McKay, 2006). However, it should be acknowledged that the disadvantage of such a test was that it was both time-consuming and costly since this test was individually administered, and it required a test administrator with near-native language proficiency; pre-training for the test administrator was also needed, further contributing to costs. Therefore, the test might not be an efficient way for local assessors to test all the students in an East Asian context characterised by large class sizes and by a lack of qualified teachers to assist with test administration.

For the speaking test, the Frog storytelling (Mayer, 1969) task was used. This task could be used to elicit a sample of students' oral production, with rich information ranging from the macro level to the micro level. Students' overall narrative compositions could be assessed by the story grammar elements; in addition, the students' productivity and complexity of language use could be measured by linguistic elements. However, similar to the administrative disadvantages of the BAS-II Verbal Comprehension, the administration, recording, transcription and scoring procedure of the *Frog* storytelling task was time-consuming. But the storytelling task might be readily implemented during daily classroom-based assessment. During the cued textbook retelling activity, participating students might come to understand that they could not creatively use the target language to compose spontaneous stories and thus forego a reliance on the exact words and phrases, as well as on the fixed plot, found in the textbook. As a result, a frequently used assessment activity commonly associated with eliciting routine information based on known information from the textbook could be productively modified to assess—and to encourage—students' spontaneous and creative language use. In sum, the use of these two tests might shed light on the future study in similar EFL contexts (i.e., in contexts where a lack of suitable instruments for

assessing listening and speaking skills is a known problem [Enever, 2011; McKay, 2006; Szpotowicz, 2012].)

In addition, in the current study, in order to interpret the descriptive results of the two assessment instruments in relation to the NECS targets, two standard-setting methods were employed to establish the cut-off scores for these assessment instruments; these scores were based on the judgement of a panel of local expert teachers. Specifically, the modified Angoff Method was used for the test assessing the examinees' receptive skills by means of multiple-choice questions; the contrasting group method was used for the test targeting the productive skills of the examinees' actual performance. The application of these two standard-setting methods might provide a possible solution for future studies or for schools and governments seeking to link numerical test scores with verbal can-do statements in the curriculum targets. Specifically, the advantage of employing the two standard-setting methods is that the panellists involved do not need to be externally invited professional experts; local teachers who meet the criteria (see section 3.4.4.2) would be capable of participation. Moreover, this method would also give local expert teachers a further opportunity to participate in the interpretation of the national curriculum targets in light of local contexts, since, during the standard-setting procedures (i.e., at the familiarisation stage), panellists would have to come to a set of agreed-upon curriculum targets that were suitable for local students.

7.4 Limitations and suggestions for the future study

In this section, three main limitations are identified with the associated suggestions for future study. These limitations correspond to the three aspects (i.e., classroom context; teachers' perspectives and students' learning outcomes) of the main research questions that have been discussed.

First, this study was intended to obtain a comprehensive picture of what was actually happening in the classroom in terms of the aural and oral assessment of young learners at two key grade levels. However, this study has focused on only four EFL teachers from two schools. The limited scope of the study may be attributed to constraints with both time and access. In addition, the classroom observations of each teacher were limited to only three lessons. Although the study aimed to observe an entire teaching unit, and the four teachers consulted had agreed that the average length of a unit was normally three lessons, in reality, one of the four teachers needed one or two extra lessons to finish the whole unit. In addition, data from each of the teachers' three lessons was, inevitably, limited in scope, and so it could not capture every kind of frequently used assessment activities, especially since some activities (e.g., dictation, English drama) were planned to be conducted less frequently (i.e., only every other week or once in a month). While these more intermittent activities were mentioned during the teacher interviews (which occurred after the classroom observations were complete), they could not be observed in class and thus the dynamic interaction and the actual implementation practice were not available for analysis.

Therefore, to aid our understanding of assessment practices and to provide further insights regarding the development of students' language skills over time in class, there is a need for future studies that (a) target a larger number of students and teachers from different contexts (e.g., rural areas) and (b) incorporate longer and more extensive classroom observations.

Second, this study aimed to explore local teachers' perspectives on curriculum implementation and their reported practices for classroom-based aural and oral assessment. To this end, the study collected substantial data from teachers, which it obtained via a survey of all the EFL teachers in the local district, as well as in-depth interviews with four of these teachers. In the future, studies could offer an even more comprehensive picture of how curriculum targets are set at the national level and how these targets are being interpreted by the local government if the studies were to also incorporate the perspectives of the following stakeholders: school administrators, local ELT coordinators and, if possible, also policymakers and curriculum designers. This additional input would help to identify the potential gaps between top-down initiatives at the national level and a bottom-up implementation at the local context level.

Third, this study employed two externally designed, individually administered assessment instruments to assess students' oral and aural skills; these tests' scores were matched to the NECS targets via standard-setting methods. The limitations associated with this part of the study were twofold. First, the two assessment instruments did not cover all of the aspects mentioned in the NECS targets regarding listening and speaking skills. For example, the frog storytelling task mainly tapped into oral production skills, and, even more specifically, into narrative skills. Other oral-related skills that were specified in the NECS, such as pronunciation or communicative skills, were not assessed because of the difficulty of finding the proper instruments and logistics reasons to do so. Therefore, the results of the two assessment instruments did not fully represent the students' overall competence in listening and speaking. Future studies might consider designing or adapting other possible assessment instruments that target other sub-skills to get more complex, richer or diverse results regarding students' aural and oral competence. Second, via the two standard-setting methods employed for these assessment instruments, the cut-off scores were identified in order to differentiate students who were at or above a certain achievement level from those who were not. However, it should be acknowledged that, in reality, there were no absolute cut-off scores for any specific test administered in the local district; rather, the cut-off scores for these instruments were established according to the local situation, following the specific procedures of one method, which served as the reference for further interpretation and other relevant information should also be taken into account at the meantime (Cizek & Bunch, 2007). Moreover, such results were obtained from only four classes of EFL students at two state primary schools in the local context, and the instruments were adapted, piloted and modified before their use in the main study phase; the results, therefore, could not be generalized to all of the other primary schools

across China. In addition, unless these instruments were carefully piloted and adapted in accordance with the local situation, the instruments also could not be used directly with young Chinese EFL learners in the primary grades. Future studies might consider conducting these assessment instruments in a large-scale way in which including the state primary schools not only from the urban cities but also from rural areas. This might help with providing more information regarding students' language abilities from different contexts in China.

7.5 Final conclusions

This thesis aims to go inside the 'black box' (i.e., the classroom) of EFL learning in a sample of state-funded Chinese primary schools in two key grade levels (Grade 4 and Grade 6) to explore and describe teachers' assessment practices and perceptions, and students' learning outcomes and their affective attitudes, as well as the relationships between these outcomes and the set targets that are articulated in the NECS. The specific focus of the EFL learning was with respect to the assessment of aural and oral language abilities. The thesis concludes that the EFL teachers tended to assess students' listening comprehension and oral production skills in terms of vocabularies and sentences, but relatively less often with respect to discourses. It also shows that current classroom teaching and assessment is still teacher centred and textbook dominated. This was found to be affected by a number of internal and external factors, but particularly contextual constraints (i.e., large class sizes and limited teaching time) and teachers' oral English proficiency. Moreover, less than half of the Grade 4 students achieved the NECS targets in both listening and speaking skills, just over half of the Grade 6 students met the expected targets in speaking skills and less than half of the Grade 6 students achieved the targets in listening skills. These findings indicate that the NECS targets were underachieved in the two sampled Chinese primary schools.

References

Allen, M. (2017). Factor Analysis: Varimax Rotation. In K. Dilbeck (Ed.), *The SAGE Encyclopedia of Communication Research Methods* (pp. 532–533). Thousand Oaks, CA: SAGE Publications. https://doi.org/10.4135/9781483381411.n191

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Bachman, L. F., & Palmer, A. S. (1996). *Language tessting in practice: Designing and developing useful language tests*. Oxford, England: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing langauge assessment and justifying their use in the real world*. Oxford, England: Oxford University Press.

Bailey, A. L. (2005). Test review:Cambridge Young Learners English (YLE)Tests. *Language Testing*, *22*(2), 242–252.

Baldauf, R., Kaplan, R., Kamwangamalu, N., & Bryant, P. (2011). Success or failure of primary second /foreign language programmes in Asia : what do the data tell us ? *Current Issues in Language Planning*, *12*(2), 309–323.

Barbour, R. (2008). *Introducing Qualitative Research: A Student Guide to the Craft of Doing Qualitative Research.* London, UK: SAGE Publications Ltd. Retrieved from http://www.sagepub.com/books/Book228282

Berk, L. (2000). *Child Development*. Needham Heights, MAA: Pearson Education Company.

Bialystok, E. (2001). *Bilingualism in development*. Cambridge: Cambridge University Press.

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–73.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising Standards Through Classroom Assessment. *Phi Delta Kappan*, *80*(2), 139–144, 146–148.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5.

Brumen, M., Cagran, B., & Rixon, S. (2009). Comparative assessment of young learners' foreign

language competence in three Eastern European countries, *5698*(March 2014), 37–41. https://doi.org/10.1080/03055690802648531

Butler, R. (1988). Enhancing and undermining intrinsic motivation: the effects of task-involving and ego-involving evaluation on insterest and performance. *British Journal of Educational Psychology*, *58*(1), 1–14.

Butler, Y. G. (2004). What Level of English Proficiency Do Elementary School Teachers Need to Attain to Teach EFL? Case Studies from Korea, Taiwan, and Japan. *TESOL Quarterly*, *38*(2), 245–278. https://doi.org/10.2307/3588380

Butler, Y. G. (2005a). Comparative perspectives towards communicative activities among elementary school teachers in South Korea, Japan and Taiwan. *Language Teaching Research*, *9*(4), 423–446. https://doi.org/10.1191/1362168805lr176oa

Butler, Y. G. (2005b). Comparative perspectives towards communicative activities among elementary school teachers in South Korea, Japan and Taiwan. *Language Teaching Research*, *4*, 423–446. https://doi.org/10.1191/1362168805lr176oa

Butler, Y. G. (2008). Variables that Influence Elementary School Students' English Performance in Japan. *The Journal of Asia TEFL*, *5*(1), 65–95.

Butler, Y. G. (2009a). How do teachers observe and evaluate elementary school students' foreign language performance? A case study from South Korea. *TESOL Quarterly*, *43*(3), 417–444.

Butler, Y. G. (2009b). Issues in the Assessment and Evaluation of English Language Education at the Elementary School Level : Implications for Policies in South. *The Journal of Asia TEFL*, *6*(2), 1–31.

Butler, Y. G. (2011). The Implementation of Communicative and Task-Based Language Teaching in the Asia-Pacific Region. *Annual Review of Applied Linguistics*, *31*, 36–57. https://doi.org/10.1017/S0267190511000122

Butler, Y. G. (2015). English language education among young learners in East Asia: A review of current research (2004–2014). *Language Teaching*, *48*(03), 303–342. https://doi.org/10.1017/S0261444815000105

Butler, Y. G., & Lee, J. (2006). On-task versus off-task self-assessments among Korean Elementary School students studying English. *Modern Language Journal*, *90*(4), 506–518.

https://doi.org/10.1111/j.1540-4781.2006.00463.x

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, *27*(1), 5–31. https://doi.org/10.1177/0265532209346370

Butler, Y. G., & Zeng, W. (2011). The roles that teachers play in paired-assessments for young learners. In D. Tsagari & I. Csépes (Eds.), *Classroom-based language assessment.* (pp. 77–92). Frankfurt: Peter Lang.

Butler, Y. G., & Zeng, W. (2014). Young foreign language learners' interactions during task-based paired assessments. *Language Assessment Quarterly*, *11*(1), 45–75. https://doi.org/10.1080/15434303.2013.869814

Carless, D. R. (2001). *Curriculum Innovation in the Primary EFL classroom : Case studies of three teachers implementing Hong Kong ' s Target-Oriented Curriculum (TOC)*. University of Warwick.

Carless, D. R. (2003). Factors in the implementation of task-based teaching in primary schools. *System*, *31*, 485–500. https://doi.org/10.1016/j.system.2003.03.002

Carreira, J. M. (2006). Motivation for Learning English as a Foreign Language in Japanese Elementary Schools. *JALT Journal*, *28*(2), 135–157.

Carroll, J. B. (1968). The psychology of language testing. In A. Davies (Ed.), *Lnaguage testing symposium. A psycholonguistic perspective* (pp. 46–69). London: Oxford University Press.

Chan, Y. (2008). Elementary school EFL teachers' beliefs and practices of multiple assessments. *Reflections on English Language Teaching*, *7*(1), 37–62. Retrieved from http://nus.edu.sg/celc/publications/RELT71/37to62-chan.pdf

Cheng, X. (2011). The "English Curriculum Standards" in China: rationales and issues. In A. Feng (Ed.), *Bilingual Education and Bilingualism : English Language Education Across Greater China.* (pp. 133–150). Bristol, GBR: Channel View Publications.

Cheung, A. C. K., & Wong, P. M. (2012). Factors affecting the implementation of curriculum reform in Hong Kong: Key findings from a large-scale survey study. *International Journal of Educational Management*, *26*(1), 39–54. https://doi.org/10.1108/09513541211194374

Chinese Ministry of Education. (2001). *全日制义务教育普通高级中学英语课程标准（实验搞）*

[English curriculum standards for full-time compulsory and senior high schools (trail version)]. Beijing: Beijing Normal University Press.

Chinese Ministry of Education. (2011). *义务教育英语课程标准 （2011版）[English curriculum standards for compulsory education (2011 version)].* Beijing: Beijing Normal University Press.

Chou, C. (2008). Exploring elementary english teachers' practical knowledge: A case study of EFL teachers in Taiwan. *Asia Pacific Education Review*, *9*(4), 529–541. https://doi.org/10.1007/BF03025668

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance on Tests.* Thousand Oaks, CA: SAGE Publications Ltd.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research Methods in Education* (Sixth Edit). Oxon: Routledge.

Council of Europe. (2009). *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) A Manual*. Strasbourg,France: Language Policy Dvision.

Cowley, J., & Glasgow, C. (1994). *The Renfrew Bus Story: Language screening by narrative recall-Examiner's manual.* Centreville, DE: The Centreville School.

Creswell, J. w., & Clark, V. L. (2007). *Designing and conducting mixed methods research*. Ohio: Pearson.

Csizer, K., & Dörnyei, Z. (2005). The internal structure of language learning motivation and its relationship with language choice and learning effort. *The Modern Language Journal*, *89*(1), 19–32.

Cummins, J. (1984). *Bilingualism and special education: issues in assess- ment and pedagogy. Multilingual matters*. Clevedon, UK.

Davison, C., & Leung, C. (2009a). Current Issues in English Language Teacher-Based Assessment.pdf. *Tesol Quarterly*, *43*(3), 393–415. https://doi.org/10.1002/j.1545-7249.2009.tb00242.x

Davison, C., & Leung, C. (2009b). Current Issues in English Language Teacher-Based Assessment. *TESOL Quarterly*, *43*(3), 393–415. https://doi.org/10.2307/27785027

Davison, C., & Leung, C. (2009c). Current Issues in English Language Teacher-Based Assessment. *Tesol Quarterly*, *43*(3), 393–415. https://doi.org/10.1002/j.1545-7249.2009.tb00242.x

Dekeyser, R. (2012). Age effects in second language learning. In S. Gass & A. Mackey (Eds.), *Handbook of second language acquisition* (pp. 442–460). London, UK: Routledge.

DeKeyser, R. (2000). The Robustness of Critical Period Effects in Second Language Acquisition. *Studies in Second Language Acquisition*, *22*, 499–533. https://doi.org/0272-2631/00

Dekeyser, R., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*(03), 413–438. https://doi.org/10.1017/S0142716410000056

Dekeyser, R., & Larson-Hall, J. (2005). What does the Critical Perid really mean? In *Handbook of bilingualism: Psycholinguistic approaches* (pp. 88–108). New York: Oxford University Press.

Deng, C., & Carless, D. (2009). The commonucativeness of activities in a task-based innovation in Guangdong, China. *Asian Journal of English Language Teaching*, *19*, 113–134.

Deng, C., & Carless, D. (2010). Examination Preparation or Effective Teaching : Conflicting Priorities in the Implementation of a Pedagogic Innovation. *Language Assessment Quaterly*, *7*(4), 285–302. https://doi.org/10.1080/15434303.2010.510899

Djigunović, J. M., Nikolov, M., & Ottó, I. (2008). A comparative study of Croatian and Hungarian EFL students. *Language Teaching Research*, *12*(3), 433–452. https://doi.org/10.1177/1362168808089926

Dockrell, J. E., Stuart, M., & King, D. (2010). Supporting early oral language skills for English language learners in inner city preschool provision. *The British Journal of Educational Psychology*, *80*(Pt 4), 497–515. https://doi.org/10.1348/000709910X493080

Elliot, C. D., Smith, P., & McCulloch, K. (1996). *British Ability Scales II*. Winsor: NFER-Nelson.

Ellis, G., Brewster, J., & Girard, D. (2002). *The Primary English Teacher's Guide* (2nd ed.). Longman.

Enever, J. (2011). *Early Language Learning in Europe*. Lond: The British Council.

Enever, J. (2018). *Policy and Politics in Global Primary English*. *Journal of Experimental Psychology: General*. Oxford, England: Oxford University Press.

Enever, J., Lindgren, E., & Ivanov, S. (2014). *Conference Proceedings from Early Language Learning : Theory and Practice 201*4. Umeå, Sweden: Dept of Language Studies, Umeå University.

Enever, J., Moon, J., & Raman, U. (2009). *Young Learner English Language Policy and Implementation: International Perspectives*. *Garnet Education Publishing*. Reading: Garnet Education Publishing. https://doi.org/10.1016/j.system.2010.06.004

Erlandson, D., Harris, E., & Skipper, B. (1993). *Doing naturalistic inquiry : A guide to methods*. Newbury Park, CA: SAGE Publications. Retrieved from http://lstmed.idm.oclc.org/login?url=https://search.ebscohost.com/login.aspx?direct=true&AuthType=sso,url,uid,cookie&db=cat03962a&AN=liv.27810116&site=eds-live&scope=site&custid=ns241346%255Cnhttp://catdir.loc.gov/catdir/enhancements/fy0655/93003651-d.html%25

Field, A. (2009). *Discovering Statistics Using SPSS 3rd Edition*. London: SAGE Publications Ltd.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *33*(349), 596–597. https://doi.org/10.1121/1.413041

Fraenkel, J. R., & Wallen, N. E. (2007). *How to design and evaluate research in eduaction*. New York. NY: McGraw-Hill Higher Education.

Garton, S., Copland, F., & Burns, A. (2011). *Investigating Global Practices in Teaching English to Young Learners*. *ELT Research Papers* (Vol. 11). London: British Council.

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2006). *Educating English Language Learners*. New York: Cambridge University Press.

Gorsuch, G. J. (2000). EFL Educational Policies and Educational Cultures: Influences on Teachers' Approval of Communicative Activities. *TESOL Quarterly*, *34*(4), 675. https://doi.org/10.2307/3587781

Goto Butler, Y., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, *27*(1), 5–31. https://doi.org/10.1177/0265532209346370

Groundwater-Smith, S., Ewing, R. A., & Jean Le Cornu, R. (2006). *Teaching: Challenges and dilemmas*. South Melbourne: Thomson.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical Evidence: A Test of the Critical-Period Hypothesis for Second-Language Acquisition. *Psychological Science*, *14*(1), 31–38. https://doi.org/10.1111/1467-9280.01415

Hambleton, R. K., & Pitoniak, M. (2006). Setting Performance Standards. In R. L. Brennan (Ed.), *Educational Measurement* (Fourth Edi, pp. 433–470). Westport, CT: Praeger.

Harlen, W. (2011). On the realtionship between assessment for formative and summative purposes. In *Assessment and Learning*. London: SAGE Publications.

Hayes, D. (2014). *Factors influencing success in teaching English in state primary schools*. London, England.

Hill, K., & McNamara, T. (2011). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing*, *29*(3), 395–420. https://doi.org/10.1177/0265532211428317

Holliday, A. (1994). *Appropriate methodology and social context*. Cambridge: Cambridge University Press.

Hu, G., & McKay, S. L. (2012). English language education in East Asia: some recent developments. *Journal of Multilingual and Multicultural Development*, *33*(February), 345–362. https://doi.org/10.1080/01434632.2012.661434

Hurtz, G. M., & Hertz, N. R. (1999). How many raters should be used for establishing cutoff scores with the angoff method? a generalizability theory study. *Educational and Psychological Measurement*, *59*(6), 885–897.

Hutcheson, G., & Sofroniou, N. (1999). *The multivariate social scientist*. London, UK: SAGE Publications Ltd.

Hutchinson, J. M., Whiteley, H. E., Smith, C. D., & Connors, L. (2003). The developmental progression of comprehension-related skills in children learning EAL. *Journal of Research in Reading*, *26*(1), 19–32. https://doi.org/10.1111/1467-9817.261003

Isbell, R., Sobol, J., Lindauer, L., & Lowrance, A. (2004). The Effects of Storytelling and Story Reading on the Oral Language Complexity and Story Comprehension of Young Children. *Early Childhood Education Journal*, *32*(3), 157–163. https://doi.org/10.1023/B:ECEJ.0000048967.94189.a3

Japanese Ministry of Education, Culture, Sports, S. and T. (2011). *The new course of studies for elementary schools.* Retrieved from http://www.mext.go.jp/.

Johnstone, R. (2009). An early start: What are the key conditions for generalized success. In J. Enever, J. Moon, & U. Raman (Eds.), *Young learner English language policy and implementation: International perspectives* (pp. 31–41). Reading,UK: Garnet Education Publishing.

Kachru, B. B. (1992). World Englishes: approaches, issues and resources. *Language Teaching*, *25*(01), 1. https://doi.org/10.1017/S0261444800006583

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53–88). Mahwah: NJ: Lawrence Erlbaum.

Kang, J. Y. (2012). How do narrative and language skills relate to each other? Investigation of young Korean EFL learners' oral narratives. *Narrative Inquiry*, *22*(2012), 307–331. https://doi.org/10.1075/ni.22.2.06yus

Karavas-Doukas, E. (1996). Using attitude scales to investigate teachers' attitudes to the communicative approach. *ELT Journal*, *50*(3), 187–198. https://doi.org/10.1093/elt/50.3.187

Khabbazbashi, N., Khalifa, H., Robinson, M., Ellis, S., & Mifsud, C. (2016). *The English Benchmarking Study in Maltese Schools: Technical Report*. Cambridge,UK.

Khabbazbashi, N., Khalifa, H., Robinson, M., Ellis, S., & Mifsud, C. (2017). Understanding language learning in Malta. *Cambridge ESOL Research Notes*, (65).

Korean Ministry of Education. (1997). *The elementary and middle school curriculum: National common basic curriculum.* Seoul.

Kormos, J., & Csizér, K. (2008). Age-related differences in the motivation of learning English as a foreign language: Attitudes, selves and motivated learning behavior., (June), 327–355. https://doi.org/10.1111/j.1467-9922.2008.00443.x

Kosuke Imai. (2017). *Quantitative social science: an introduction*. Princeton, NJ: Princeton University Press.

Kumaravadivelu, B. (2001). Toward a Postmethod Pedagogy. *Tesol Quarterly*, *35*(4), 537–560.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

Lantolf, J. P. (2000). *Sociocultural Theory and Second Language Learning.* Oxford, England: Oxford University Press.

Lasagabaster, D. (2011). English achievement and student motivation in CLIL and EFL settings. *Innovation in Language Learning and Teaching*, *5*(1), 3–18. https://doi.org/10.1080/17501229.2010.519030

Lenneberg, E. H. (1967). *Biological foundations of language*. New York: Wiley.

Leung, C. (2007). Dynamic assessment: assessment for and as teaching? *Language Assessment Quarterly*, *4*, 257–278. https://doi.org/10.1080/15434300701481127

Li, M., & Baldauf, R. (2011). Beyond the Curriculum：A Chinese Example of Issues Constraining Effective English Language Teaching. *TESOL Quarterly*, *45*(4), 793–803. https://doi.org/10.5054/tq.201

Liddicoat, A. J. (1997). *Communicating in LOTE. Writing and Oral Interaction*. Canberra: Modern Language Teachers' Association of the Australian Captical Territory.

Lindgren, E., & Enever, J. (2017). Mixed methods in early language language learning research-examining complexity. In E. Lindgren & J. Enever (Eds.), *Early language learning: complexity and mixed methods* (pp. 22–28). Bristol, GBR: Multilingual Matters.

Livingston, S. A., & Zieky, M. J. (1982). *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests*. Princeton,NJ: Educational Testing Service.

Long, M. (2005). Problems with supposed counter-evidence to the. *International Review of Applied Linguistics in Language Teaching (IRAL)*, *43*(2005), 287–317.

Lopriore, L., & Mihaljević Djigunović, J. (2009). Attitudinal aspects of early EFL learning. In *UPRT 2009: Empirical research in English applied linguistics.* Pecs, Hungary: Lingua Franca Csoport.

Luo, S. (2014). Task-Based Language Teaching and Assessment in Chinese Primary and Secondary Schools. In *English Language Education and Assessment* (pp. 205–220). Singapore: Springer.

MacIntyre, P. D., Baker, S. C., Clément, R., & Donovan, L. A. (2002). Sex and age effects on

willingness to communicate, anxiety, perceived competence, and L2 motivation among junior high school French immersion students. *Language Learning*, *52*(3), 537–564. https://doi.org/10.1111/1467-9922.00194

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design*. Routledge.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* (3rd Editio). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2005). A unified model of language development. In J. . Kroll & A. M. . De Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 49–67). Oxford, England: Oxford University Press.

Mayer, M. (1969). *Frog, where are you?* New York; NY: Dial Press.

Mayer, M. (1974). *Frog goes to dinner.* New York; NY: Dial Press.

McKay, P. (2006). *Assessing Young Language Learners*. Cambridge,UK: Cambridge University Press.

Mechelli, A., Crinion, J. T., Noppeney, U., O'Doherty, J., Ashburner, J., Frackowiak, R. S., & Price, C. J. (2004). Structural plasticity in the bilingual brain. Proficiency in a second language and age at acquisition affect grey-matter density. *Nature*, *431*(7010), 2–3. https://doi.org/10.1038/nature03017

Mehan, H. (1979). *Learning lessons: Social organization in the classsroom*. Cambridge, MA: Harvard University Press.

Merriam, S. B. (1998). *Case study research in education : a qualitative approach*. San Francisco: Jossey-Bass Inc.,U.S.

Mihaljević Djigunović, J. (2009). Individual differences in early language programmes. In M. Nikolov (Ed.), *The age factor and early language learning* (pp. 199–226). Berlin,Germany: Mouton de Gruyter.

Mihaljević Djigunović, J. (2015). Individual differences among young EFL learners: age- or proficiency-related? A look from the affective learner factors perspective. In J. Mihaljević Djigunović & M. Medved Krajnovic (Eds.), *Early learning and teaching of English: New dynamics of primary English* (pp. 10–36). Bristol, GBR: Multilingual Matters.

Mihaljević Djigunović, J., & Legac, V. (2008). Foreign Language Anxiety and Listening Comprehension of Monolingual and Bilingual EFL Learners. *Sraz*, *53*, 327–347.

Mihaljević Djigunović, J., & Letica Krevelj, S. (2010). Instructed early SLA: Development of attitudes. *SRAZ*, *54*.

Mihaljević Djigunović, J., & Lopriore, L. (2011). The learner: Do individual differences matter? In J. Enever (Ed.), *ELLiE: Early language learning in Europe* (pp. 29–45). London, UK: The British Council.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (Second). SAGE Publications.

Muñoz, C. (2006). The effects of age on foreign language learning: the BAF project. In C. Muñoz (Ed.), *Age and the rate of foreign langauge learning*. Bristol: Multilingual Matters.

Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, *35*(4), 463–482. https://doi.org/10.1093/applin/amu024

Muñoz, C. (2017). Tracing Trajectories of Young Learners: Ten Years of School English Learning. *Annual Review of Applied Linguistics*, *37*(2017), 168–184. https://doi.org/10.1017/S0267190517000095

Murphy, V. (2014). *Second language learning in the early school years: Trends and contexts.* Oxford, England: Oxford University Press.

Nemoto, A. K. (2018). Getting Ready for 2020: Changes and Challenges for English Education in Public Primary Schools in Japan. *JALT Publications*, *The Language Teacher* (42/4).

Nicolopoulou, A. (2008). The elementary forms of narrative coherence in young children's storytelling. *Narrative Inquiry*, *18*(2), 299–325. https://doi.org/10.1075/ni.18.1.07nic

Nikolov, M., & Curtan, H. (2000). *An early start: Young learners and modern languages in Europe and beyond.* Strasbourg, France: Council of Europe.

Nikolov, M., & Mihaljević Djigunović, J. (2006). Recent Research on Age, Second Language Acquisition, and Early Foreign Language Learning. *Annual Review of Applied Linguistics*, *26*, 234–260.

Nikolov, M., & Mihaljević Djigunović, J. (2011). All Shades of Every Color: An Overview of Early Teaching and Learning of Foreign Languages. *Annual Review of Applied Linguistics*, *31*, 95–

119. https://doi.org/doi:10.1017/S0267190511000183

Nunan, D. (2003). The Impact of English as a Global Language on Educational Policies and Practices in the Asia-Pacific Region. *Tesol Quarterly*, *37*(4), 589–613. https://doi.org/10.2307/3588214

Patri, M. (2002). The influence of peer feedback on self-and peer-assessment of oral skills. *Language Testing*, *19*(2), 109–131. https://doi.org/10.1191/0265532202lt224oa

Patton, M. Q. (2014). *Qualitative research & evaluation methods: integrating theory and practice*. London: SAGE Publications.

Petersen, D. B., Gillam, S. L., & Gillam, R. B. (2008). Emerging Procedures in Narrative Assessment: The Index of Narrative Complexity. *Topics in Language Disorders*, *28*(2), 115–130.

Petersen, D., Gillam, S. L., Spencer, T., & Gillam, R. (2010). The effects of literate narrative intervention on children with neurologically based language impairments: an early stage study. *Journal of Speech, Language, and Hearing Research : JSLHR*, *53*(4), 961–981. https://doi.org/10.1044/1092-4388(2009/09-0001)

Prapaisit De Segovia, L., & Hardison, D. M. (2008). Implementing education reform: EFL teachers' perspectives. *ELT Journal*, *63*(2), 154–162. https://doi.org/10.1093/elt/ccn024

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, *2*(1), 13–43. https://doi.org/10.1159/000310933

Puckett, M., & Black, J. (2000). *Authentic assessment of the young child: celebrating development and learning*. Merill.

Punch, K. F., & Oancea, A. (2014). *Introduction to research methods in education* (2nd ed.). London, England: SAGE Publications.

Rea-Dickins, P. (2000). Assessment in early years language learning contexts. *Language Testing*, *2*, 115–122. https://doi.org/10.1177/026553220101800407

Rea-Dickins, P. (2001). Mirror, mirror on the wall: identifying processes of classroom assessment. *Language Testing*, *18*(4), 429–462. https://doi.org/10.1191/026553201682430120

Rea-Dickins, P., & Gardner, S. (2000). Snares and silver bullets: disentangling the construct of formative assessment. *Language Testing*, *17*(2), 215–243.

Rea-Dickins, P., & Rixon, S. (1999). Assessment of young learners' English: Reasons and means. In *Young learners of English: Some research perspectives* (pp. 89–101). Harlow, Essex: Longman.

Rixon, S. (2013). *British Council Survey of Policy and Practice in Primary English Language Teaching Worldwide*. Retrieved from http://www.teachingenglish.org.uk/sites/teacheng/files/D120 Survey of Teachers to YLs_FINAL_Med_res_online.pdf

Scarino, A., Vale, D., McKay, P., & Clarks, J. . (1988). *The Australian Language Levels Guidelines (Books 1-4)*. Canberra: Curriculum Development Centre.

Schumann, J. (1997). *The neurobiology of affect in language*. Malden, MA: Blackwell.

Sen, R., & Blatchford, P. (2001). Readings in a Second Language: Factors associated with progress. *Educational Psychology*, *21*(2), 189–202. https://doi.org/10.1080/01443410020043887

Singleton, D. (2014). Apt to change: The problematic of language awareness and language aptitude in age-related research. *Studies in Second Language Learning and Teaching*, *3*(3), 557–571. https://doi.org/10.14746/ssllt.2014.4.3.9

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.

Spada, N., & Fröhlich, M. (1995). *Communicative orientation of language teaching observation scheme coding conventions and applications*. Sydney, Australia: Macquarie University, NCELTR.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R.O.Freedle (Ed.), *New directions in discourse processing* (pp. 53–120). NJ: Ablex.

Su, Y. C. (2006). EFL teachers' perceptions of English language policy at the elementary level in Taiwan. *Educational Studies*, *32*(3), 265–283. https://doi.org/10.1080/03055690600631218

Szpotowicz, M. (2012). Researching oral production skills of young learners. *CEPS Journal*, *2*(3), 141–166.

Szpotowicz, M., & Lindgren, E. (2011). Language achievements: a longitudinal perspective. In Janet Enever (Ed.), *ELLiE: Early Language Learning in Europe*. London: The British Council.

Szpotowicz, M., Mihaljević Djigunović, J., & Enever, J. (2009). Early language learning in Europe: A

multinational, longitudinal study. In Janet Enever & U. Raman (Eds.), *Young learner English language policy and implementation: International perspectives* (pp. 141–147). Reading,UK: Garnet Publishing.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (6th ed.). Bonston: Peason/Allyn & Bacon.

Tannenbaum, R. J., & Wylie, E. C. (2008). TOEFL iBT Research Report Linking English-Language Test Scores Onto the Common European Framework of Reference : An Application of Standard-Setting Methodology, (June).

Topping, K. J. (2009). Peer assessment. *Theory into Practice*, *48*(1), 20–27. https://doi.org/10.1080/00405840802577569

Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal*, *27*(5), 615–631. https://doi.org/10.1080/01411920120095780

Tunstall, P., & Gipps, C. (1996). Teacher Feedback to Young Children in Formative Assessment : A Typology. *British Educational Research Journal*, *22*(4), 389–404.

Wang, H. (2006). *An implementation study of the English aas a foreign language curriculum policues in the Chinese tertiary context*. Queen's University.

Wang, Q. (2007). *Primary EFL in China : Teachers ' Perceptions and Practices with Regard to Learner-centredness*. University of Warwick

Wang, Q. (2009). Primary English in China: policy, curriculum and implementation. In M. Nikolov (Ed.), *Studies on Language Acquisition : The Age Factor and Early Language Learning.* Berlin, DEU: Mouton de Gruyter.

Wang, X. (2017). A Chinese EFL Teacher's Classroom Assessment Practices. *Language Assessment Quarterly*, *14*(4), 312–327. https://doi.org/10.1080/15434303.2017.1393819

Westerveld, M. F., & Gillon, G. T. (2008). Oral narrative intervention for children with mixed reading disability. *Child Language Teaching and Therapy*, *24*(1), 31–54. https://doi.org/10.1177/0265659007084567

White, E. J., Hutka, S. A., Williams, L. J., & Moreno, S. (2013). Learning, neural plasticity and sensitive periods : implications for language acquisition, music training and transfer across the

lifespan. *Frontiers in Systems Neuroscience*, *7*(November), 1–18. https://doi.org/10.3389/fnsys.2013.00090

Wiliam, D., & Thompson, M. (2007). Integrating assessment with learning: What will it take to make it work? *The Future of Assessment: Shaping Teaching and Learning*, 53–82. Retrieved from http://books.google.co.uk/books/about/The_future_of_assessment.html?id=bW0OAQAAMAAJ&pgis=1

Wiliam, D., & Thompson, M. (2007). Integrating assessment with instruction: What will it take to make it work? In C. A. Dwyer (Ed.), *The furture of assessment: Shaping teaching and learning* (pp. 53–82). Mahwah, NJ: Erlbaum.

Yan, J. (2014). *Exploring Classroom-Based Assessment Activities and Learner Roles in Two Tertiary Institutions in Guangdong , China*. The University of Hong Kong.

Zhang, E. Y., & Adamson, B. (2007). Implementing language policy: Lessons from primary school English. In A. Feng (Ed.), *Bilingual education in China: Practices, policies and concepts.* (pp. 166–181). Clevedon, UK: Multilingual Matters.

Zhang, X., & Head, K. (2009). Dealing with learner reticence in the speaking class. *ELT Journal*, *64*(1), 1–9. https://doi.org/10.1093/elt/ccp018

Appendices

Appendix 1 Level descriptors of language skills in the NECS

Level 1 descriptors of language skills in the NECS

	Skills	Targets
LEVEL 1	Listening	1. Can recognise pictures or real objects based on heard vocabulary and sentences.
		2. Can understand and react to simple classroom instructions.
		3. Can take action based on directions. (e.g. pointing pictures, colouring, drawing and acting).
		4. Can understand simple stories with clues of pictures and actions.
	Speaking	1. Can imitate pronunciation based on the tapes.
		2. Can greet in simple sentences.
		3. Can exchange personal information (e.g. name and age).
		4. Can express simple emotions (e.g. like and dislike).
		5. Can guess the meaning and figure out the words based on acting.
		6. Can sing 15 children's songs and rhymes.
		7. Can produce words and phrases based on given pictures and text.
	Reading	1. Can recognize words printed with pictures.
		2. Can recognize and read aloud learned words with the ability to match words with their referents.
		3. Can read and understand short picture stories.
	Writing	1. Can write letters and words correctly.
		2. Can write words and sentences with reference to exemplars.

Level 2 descriptors of language skills in NECS

	Skills	Targets
LEVEL 2	Listening	1. Can listen and understand conversation and recordings with clues of pictures, images, gesture.
		2. Can listen and understand picture stories.
		3. Can understand simple questions in class
		4. Can understand common instructions and directions and response properly.
	Speaking	1. Can articulate English clearly with the appropriate intonation.
		2. Can have simple conversation based on one's familiar personal or family situation.
		3. Can use daily expressions (e.g. greetings, express one's thanks, apologize) and express simple emotions (e.g. like and dislike).
		4. Can narrate based on daily topics.
		5. Can describe or narrate simple stories with help of teachers and clues of pictures.
	Reading	1. Can recognize learned words and phrases
		2. Can read simple words following spelling rules.
		3. Can read and understand instruction or directions in the textbooks.
		4. Can read and understand simple information on greeting cards.
		5. Can read simple stories or essays with the help of pictures and can develop the habit of reading in phrases
		6. Can read aloud accurately the learned texts or stories.
	Writing	1. Can correctly use upper- and lower-case letters and punctuation marks.
		2. Can write simple greetings.
		3. Can write simple and short sentences with clues from pictures, words and phrases, and exemplary sentences.

Appendix 2 Teacher questionnaire

Chinese primary school EFL teachers' perceptions of and attitudes towards the National English Curriculum Standards and its implementation in listening and speaking.

Instruction:

This questionnaire is designed for Chinese EFL teachers who teach in a state primary school. It aims to explore your perceptions of and attitudes towards the new National English Curriculum Standards (2011 version) and its implementation in your classroom, especially regarding the listening and speaking aspects of the standards.

There are three sections in this survey. Section A asks about your personal background information as a primary school teacher, Section B enquiries about your perceptions on curriculum targets and your attitudes towards curriculum implementation, Section C asks about your practices of assessment of young EFL learners.

Section A. Background information

Please tick an appropriate answer or write out your answer in words as required.

Gender	☐ Male	☐ Female
Age	☐ 21-29 ☐ 40-49	☐ 30-39 ☐ over 50
Educational qualifications	☐ B.A. in English ☐ M.A/M.Ed. in English ☐ College diploma in English ☐ Senior high school graduate	☐ B.A. in other subjects ☐ M.A. in other subjects ☐ College diploma in other subjects
Teaching experience	☐ 1-4 years ☐ 10-14 years ☐ Over 20 years	☐ 5-9 years ☐ 15-19 years
Current teaching grade	☐ Grade 3 ☐ Grade 5	☐ Grade 4 ☐ Grade 6
Teaching workload (class size, teaching hours)	Total number of classes you teach_____ Total number of students _____	Number of hours you teach per week _____ Average number of students per class _____

Section B. Perceptions of curriculum targets and its implementation

B1. Attitudes towards curriculum implementation

Please indicate the degree to which you agree with each of the statements in the following sections by circling the following response scale:

5 = Strongly Agree (SA)

4 = Agree (A)

3 = Neutral (N) (Neither agree nor disagree)

2 = Disagree (D)

1 = Strongly Disagree (SD)

	SA	A	N	D	SD
I could use English to discuss with students.	5	4	3	2	1
I could use English to raise high standard questions.	5	4	3	2	1
The content of the textbook provides me detailed learning objectives in teaching listening and speaking.	5	4	3	2	1
The teacher's book provides me enough guidelines on how to teach listening and speaking.	5	4	3	2	1
I have other resources when teaching listening and speaking.	5	4	3	2	1
The local final exam influences me what to teach.	5	4	3	2	1
The local final exam influences me how to teach.	5	4	3	2	1
Junior high school entry exam influences my teaching focus in listening and speaking.	5	4	3	2	1
NECS provides me detailed learning objectives in listening and speaking.	5	4	3	2	1
The suggestions from the district-level ELT coordinator contribute to the improvement of my English teaching ability.	5	4	3	2	1
Participating in the national in-service training could help me to better understand the NECS.	5	4	3	2	1
I can use English to give detailed feedback.	5	4	3	2	1
I'm confident in my English oral competence.	5	4	3	2	1
I'm confident in my English aural competence.	5	4	3	2	1
The large class size makes it challenging for me to teach speaking and listening.	5	4	3	2	1
I have moderate teaching workloads. (reversed one)	5	4	3	2	1
I have don't enough time to let students practice their oral English in my class.	5	4	3	2	1

B2. Perceptions of curriculum targets

Note: If your current students are in Grade 3 and 4, please answer Q1; if you current students are in Grade 5 and 6, please answer Q2.

Q1. To what degree do you think the following targets could be achieved by your current students by the end of Grade 4 (i.e. after two-year foreign language learning in the instructed context)?

1 = Not appropriate

2 = Approximately 25% of students in your class could achieve this target.

3 = Approximately 50% of primary school students could achieve this target.

4 = Approximately 75% of primary school students could achieve this target.

5 = Approximately 100% of primary school students could achieve this target.

Skills	Targets	N/A	25%	50%	75%	100%
Listening	1. Student can recognize pictures or real objects based on heard vocabulary and sentences.	1	2	3	4	5
(Level 1)	2. Student can understand and react to simple classroom instruction.	1	2	3	4	5
	3. Student can take action based on heard directions (e.g. pointing pictures, colouring, drawing and acting).	1	2	3	4	5
	4. Student can listen and understand simple stories with clues of pictures and actions.	1	2	3	4	5
Speaking	1. Student can imitate pronunciations based on the tapes.	1	2	3	4	5
(Level 1)	2. Student can greet each other using basic expressions.	1	2	3	4	5
	3. Student can exchange personal information in a basic phrase (e.g. name and age).	1	2	3	4	5
	4. Student can express simple emotions (e.g. like and dislike)	1	2	3	4	5
	5. Student can guess the meaning and figure out the words based on acting.	1	2	3	4	5
	6. Student can produce words and phrases based on given pictures and text.	1	2	3	4	5

Q2. To what degree do you think the following targets could be achieved by your current students by the end of ***Grade 6*** (i.e. after four years foreign language learning in the instructed context)?

Skills	Targets	N/A	25%	50%	75%	100%
Listening (Level 2)	1. Student can listen and understand conversation and recordings with clues of pictures, images, gesture.	1	2	3	4	5
	2. Student can understand simple stories supported by pictures.	1	2	3	4	5
	3. Student can understand basic questions in the context of class.	1	2	3	4	5
	4. Student can understand common instructions and directions and response properly.	1	2	3	4	5
Speaking (Level 2)	1. Student can articulate English clearly with the appropriate intonation.	1	2	3	4	5
	2. Student can have simple conversations based on one's familiar personal or family situation.	1	2	3	4	5
	3. Student can use daily expressions (e.g. greetings, express one's thanks, apologize) and express simple emotions (e.g. like and dislike).	1	2	3	4	5
	4. Student can narrate based on daily topics.	1	2	3	4	5
	5. Student can describe or narrate simple stories with the help of teachers and the clues of pictures.	1	2	3	4	5

Section C. Attitudes and practices of assessment of young EFL learners

Please indicate the frequency of your assessment practices with each of the statements in the following sections by circling the following response scale:

5 = Always: once a week (A); 4 = Frequently: once a fortnight (F); 3 = Some times: 2-3 times a term (ST); 2 = Rarely: once a term (R); 1 = Never (N)

	A	F	ST	R	N
Q1. Agent of assessment practices					
Student self-assessment	5	4	3	2	1
Peer assessment	5	4	3	2	1
Teacher led assessment	5	4	3	2	1
Q2. Focus of assessment practices					
Spelling	5	4	3	2	1
Pronunciation and intonation	5	4	3	2	1
Lexis	5	4	3	2	1
Grammar	5	4	3	2	1
Functional language	5	4	3	2	1
Content and topics	5	4	3	2	1
Affective attitudes (motivation & confidence)	5	4	3	2	1
Q3. What kind of exercises/activities do you use in assessing speaking and listening?					
Read aloud	5	4	3	2	1
Listen and imitate	5	4	3	2	1
Q&A led by the teacher	5	4	3	2	1
Listen and answer	5	4	3	2	1
Listen and act	5	4	3	2	1
Recite text/dialogues	5	4	3	2	1
Textbook role play	5	4	3	2	1
Textbook retelling	5	4	3	2	1
Q&A led by the group/pairs	5	4	3	2	1
Make up conversations	5	4	3	2	1
Dictation	5	4	3	2	1
Make up sentences	5	4	3	2	1
Oral gap filling	5	4	3	2	1
Oral translation	5	4	3	2	1
Listening comprehension (in a form of paper-and-pencil exercises)	5	4	3	2	1
Picture description	5	4	3	2	1
Storytelling	5	4	3	2	1

Appendix 3 Student questionnaire

Chinese EFL young learner's perceptions and attitudes towards EFL learning at primary level

Instruction:

The purpose of this questionnaire is to see what you think about EFL learning at school. There are no right or wrong answers to any of the items on the questionnaire. Please answer as honestly as you can based on how you really feel, not on how you think most people feel or how you think you ought to feel. Please tick in the appropriate box.

Section A. Background information

Please tick an appropriate answer or write out your answer in words as required.

1. Gender	☐ Male	☐ Female
2. Age	_____	
3. Starting age of EFL learning	_____ years old	
4. Current grade	☐ Grade 3	☐ Grade 4
	☐ Grade 5	☐ Grade 6

Section B. Views on assessment activities

Q. How much do you think these activities help you to learn English better? Please indicate your preference with each of the statements in the following sections by circling the following response scale:

☺☺= Very useful (VU+)
☺ = Useful (U+)
😐= Neutral (N) (Neither useful nor useless)
☹ = Useless (U-)
☹☹ = Very useless (VU-)

	VU+	**U+**	**N**	**U-**	**VU-**
Read aloud	☺☺	☺	😐	☹	☹☹
Listen and imitate	☺☺	☺	😐	☹	☹☹
Q&A led by the teacher	☺☺	☺	😐	☹	☹☹
Listen and answer	☺☺	☺	😐	☹	☹☹
Listen and act	☺☺	☺	😐	☹	☹☹
Recite text/dialogues	☺☺	☺	😐	☹	☹☹
Textbook role play	☺☺	☺	😐	☹	☹☹
Textbook retelling	☺☺	☺	😐	☹	☹☹
Q&A led by the group/pairs	☺☺	☺	😐	☹	☹☹
Make up conversations	☺☺	☺	😐	☹	☹☹
Dictation	☺☺	☺	😐	☹	☹☹
Make up sentences	☺☺	☺	😐	☹	☹☹
Oral gap filling	☺☺	☺	😐	☹	☹☹
Oral translation	☺☺	☺	😐	☹	☹☹
Listening comprehension (in a form of paper-and-pencil exercises)	☺☺	☺	😐	☹	☹☹
Picture description	☺☺	☺	😐	☹	☹☹
Storytelling	☺☺	☺	😐	☹	☹☹

Section C Attitudes towards EFL learning (motivation, anxiety and confidence)

Please indicate the degree to which you agree with each of the statements in the following sections by circling the following response scale:

☺☺= Strongly Agree (SA); ☺ = Agree (A); 😐= Natural (N) (Neither agree nor disagree); ☹ = Disagree (D); ☹☹ = Strongly Disagree (SD)

	SA	A	N	D	SD
Anxiety					
I'm nervous when I hear new words or unfamiliar words in English.	☺☺	☺	😐	☹	☹☹
I'm nervous if somebody speaks English too quickly.	☺☺	☺	😐	☹	☹☹
I get nervous when the teacher asks me to respond to her/him in English.	☺☺	☺	😐	☹	☹☹
I can feel my heart pounding when I am going to be called on in English class.	☺☺	☺	😐	☹	☹☹
I worry if I can acquire new words and expressions in English.	☺☺	☺	😐	☹	☹☹
I feel badly when I cannot answer the teacher's questions in English.	☺☺	☺	😐	☹	☹☹
I worry when I do not understand what the teacher is saying in English.	☺☺	☺	😐	☹	☹☹
Confidence					
I think that I speak English better than others in my class.					
I am good at learning English.	☺☺	☺	😐	☹	☹☹
I think I can be fluent in spoken English.	☺☺	☺	😐	☹	☹☹
I am confident when speaking English in class.	☺☺	☺	😐	☹	☹☹
English class is easy for me.	☺☺	☺	😐	☹	☹☹
I am doing well in my English class.	☺☺	☺	😐	☹	☹☹
I enjoy English classes.	☺☺	☺	😐	☹	☹☹
Intrinsic motivation					
I like studying English.	☺☺	☺	😐	☹	☹☹
I want to be able to speak both English and Chinese well.	☺☺	☺	😐	☹	☹☹
I want to know more about foreign countries.	☺☺	☺	😐	☹	☹☹
It will make me feel good to speak English.	☺☺	☺	😐	☹	☹☹
It is good for Chinese to be able to speak both English and Chinese.	☺☺	☺	😐	☹	☹☹
Learning English would make me feel better about myself in school.	☺☺	☺	😐	☹	☹☹
I like English TV programs, videos, and movies.	☺☺	☺	😐	☹	☹☹
Extrinsic motivation					
I think that English will be useful for me in the future.	☺☺	☺	😐	☹	☹☹
I want to be fluent in English because speaking English is useful.	☺☺	☺	😐	☹	☹☹
I want to get good scores in English tests.	☺☺	☺	😐	☹	☹☹
I want to be good at playing English video games.	☺☺	☺	😐	☹	☹☹
English is an important language to learn because it is widely spoken.	☺☺	☺	😐	☹	☹☹
I think knowing both English and Chinese helps one to get a good job in the future.	☺☺	☺	😐	☹	☹☹
I want to make many foreign friends.	☺☺	☺	😐	☹	☹☹

Appendix 4 Teacher interview schedule

Before the interview, the following issues will be clarified:

- Purpose of the interview
- Confidentially
- Format and length of the interview
- Recording issues: ask for permission of the recorder (as stated in the information sheet and consent form)

Proposed interview questions

Warming up questions:

1. How do you get into teaching?
2. What do you enjoy most about your teaching?

Section 1 Background information

1. Can you tell something about yourself, such as your educational qualification and teaching experience?
2. What grades are you teaching? How many classes are you teaching?
3. On average, how many English lessons does each class have per week? And how many students are there in a class?

Section 2 From curriculum standards perspective

1. What's your general understanding of the NECS? (the main objectives, the practicality and appropriateness)
2. What's your interpretation of the language skills targets in listening and speaking that specified in the NECS? Is it feasible for your current students? If not, why not?
3. What are the most challenging aspects that you face regarding implementing the NECS, specially, achieving the targets that specified in the NECS?

Section 3 From teaching perspective

1. What's your typical teaching procedure, especially in listening and speaking teaching? (tasks/activities, teaching materials and teaching approaches)
2. What's your teaching approach in teaching listening and speaking?
3. What kind of activities/tasks do you use in your listening and speaking teaching? And how and how often do you use it in your class?
4. How do you usually conduct your teaching to achieve the targets stated in the NECS?

Section 4 From the assessment perspective

1. Do you know about the implementation of multiple assessments (formative and summative assessments), which has been advocated in the NECS?
2. Have you got any training on the implementation of multiple assessments?
3. What do you think are the most important components of an effective formative assessment task?
4. What's your assessment procedures in class?
5. What activities do you use when assessing listening and speaking skills formatively? What skills do you focus or measure in these activities? And which activities do you think are the most effective and suitable for use on your current students in assessing and improving their listening and speaking skills?
6. How do you deal with students' error during the classroom assessments? What kind of comments/feedback do you usually give to your students formatively and summatively?
7. Do you use peer assessment in the classroom? if so, how do you conduct the peer assessment activities? Do you think the peer assessment could inform you about the progress of the learning in related with the teaching objectives?

Appendix 5 Student interview schedule

Before the interview, the following issues will be clarified:

- Purpose of the interview
- Confidentially
- Format and length of the interview
- Recording issues: ask for permission of the recorder (as stated in the information sheet and consent form)

Proposed interview questions

1. Do you like this week's lessons?
2. What do you think your teacher want you to learn during the three lessons in this week?
3. Which part did you like/dislike most, and why?
4. What kind of classroom activities do you usually enjoy most? Why?
5. Which kind of activities do you think can help you develop you oral and aural skills?
6. What kind of comments/feedback on your performance do you usually get from teachers during the classroom activities, on your homework and on your final examination? Based on teacher's feedback, do you know what you can do and what you can't do (strong and weak points) in EFL learning?
7. Based on teacher's feedback, do you know what you can do and what you can't do in terms of the EFL learning? Regarding the four language skills (i.e. listening, speaking, reading and writing), which aspect is your strong area, and which one is your weak area?
8. What do you think of your assignments? Are they difficult and are they helpful for your EFL learning? Do teachers usually provide any feedback on it?
9. What do you think of the peer assessment? Is it helpful or not?

Appendix 6 The adapted COLT scheme

Title of the lesson: _____ Textbook and page number: _____

School_____ Grade_____ Class_____ Teacher_____ Lesson (min.) _____ Visit No. _____ Date_____

Time	Assessment Activities	Assessment activities features													Assessment practice dimensions								
		Organisation			Content			Skills			Materials				Agent		Target			Actual focus		
		Class	Group	Individual	Form	Meaning	Combined	Listening	Speaking	L&S	Text Minimal	Text Extended	Audio	Visual	Teacher	Peers	Class	Group	Individual	Form	Meaning	Combined
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23

Appendix 7 British Ability Scale II: Verbal Comprehension record booklet

Student ID No._____ Age _____Class_____Grade_____ School_____ Date_____

Directions: The child points to pictures or manipulates objects in response to oral instructions from the administrator.

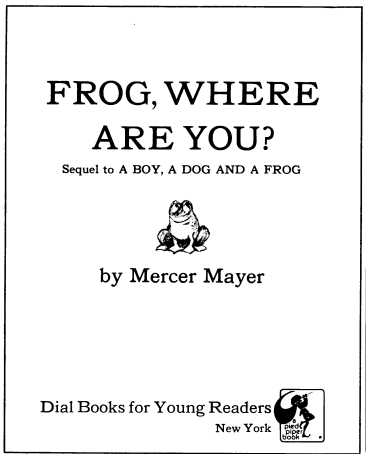
Materials: Picture of Teddy Bear; Box of Toys; Inset Tray; Nine coloured plastics shapes; Early Years Core I Booklet

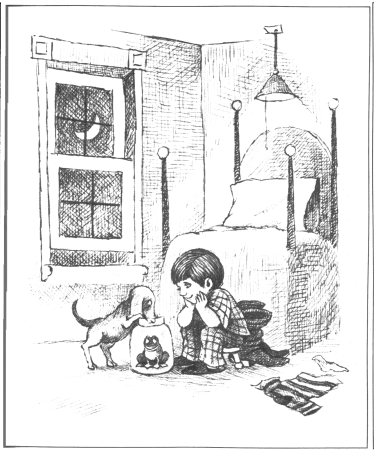
Item	Correct Response	Score
	Teddy Bear	
1	Shows bear's eyes	
2	Shows bear's legs	
3	Shows bear's arms	
4	Shows bear's ears	
5	Shows bear's mouth	
6	Gives horse	
	Toys	
7	Gives car	
8	Gives watch	
9	Puts horse in box	
10	Gives watch	
11	Gives pencil	
12	Gives cat, dog, horse	
13	Puts tree behind back	
14	Pushes van toward you	
15	Stands child on bridge	
16	Puts car under bridge	
17	Picks up car after tree is down	
18	Makes children face each other	
19	Puts little tree beside big house	
20	Gives little house then van	
21	Puts car between houses	

22	Gives car while touching van	
23	Puts child between trees	
Coloured Shapes (deleted for main study)		
24	Puts circle beyond square	
25	Gives blue square and blue triangle	
26	Picks up blue and yellow circles	
27	Gives red circle and red triangle	
28	Gives all red shapes	
29	Gives all red shapes and blue and yellow circles	
30	Gives blue and yellow circles and blue and yellow triangles	
Pictures		
31	Kevin wanted a skateboard for his birthday but got a football instead. (4)	
32	Show me the picture where the cat was followed by the dog. (2)	
33	When Lisa's father works in the garden, he puts on his old coat and boots, but never wears a hat. (4)	
34	If Carol wears her coat, she doesn't do up the buttons, but she always wears her hat. (2)	
35	The boy she waved to was sitting down. (3)	
36	Mother said to Khalid, "You should have put your jacket on this morning". (1)	
37	When Barry's dad came home, Barry forgot to warn him about the broken chair. Then he was really sorry about what happened. What did happen? (4)	
38	Just as Karen was leaving for school it started to rain. Her mother told her not to get wet and Karen did as she was told. (2)	
39	Jane's father was pleased that he had told the children never to run into the road. (2)	
40	Mother said to Ben and Rachel, "Even though it's a nice day, I want you to stay indoors to help me with all this washing up." The children got it finished really quickly. What happened then? (3)	

Appendix 8 Storytelling task

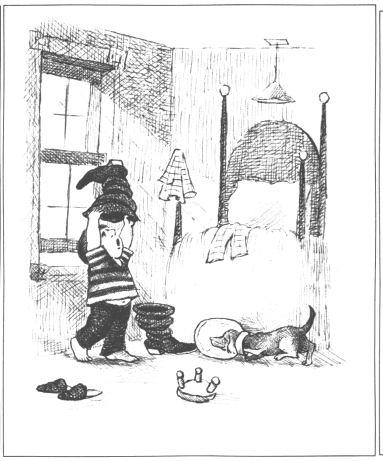
Directions: This is a story about a boy, a frog and a dog. Please look through the pictures and then tell a story.

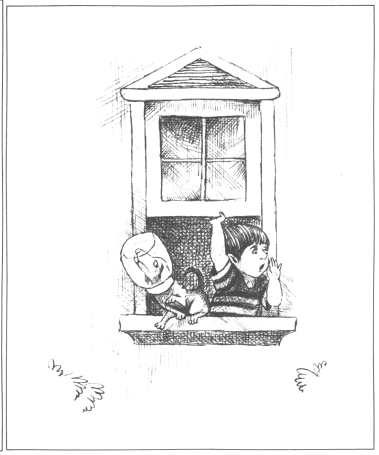


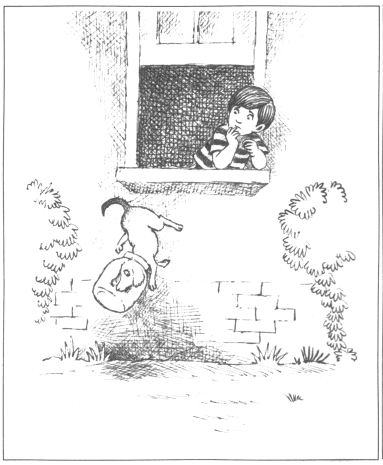


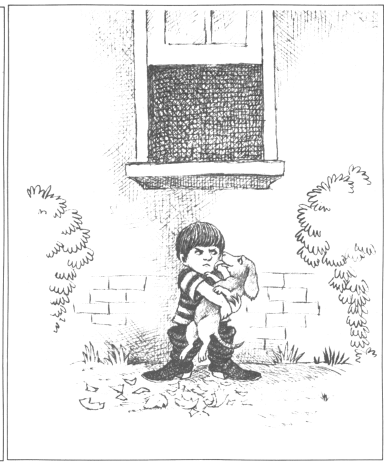






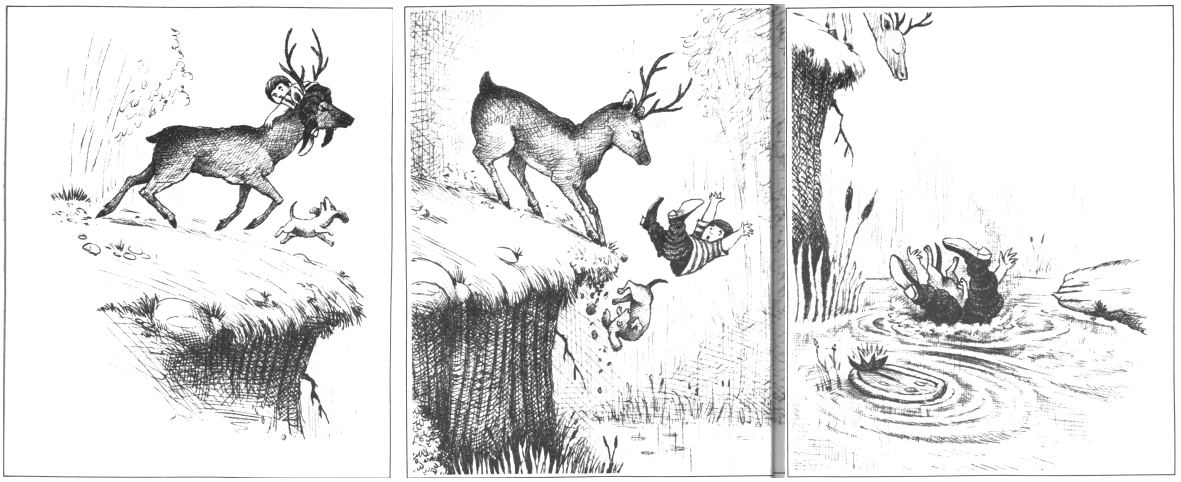
















Appendix 9 Index of Narrative Complexity Story Coding Form

Index of Narrative Complexity Story Coding Form				
Narrative Element	**0 POINTS**	**1 POINT**	**2 POINTS**	**3 POINTS**
Character *A character is any reference to the subject of a clause in a narrative.*	**No** main character is included, or only ambiguous pronouns are used. **Examples** *a) <u>They</u> were walking.* *b) <u>He</u> was walking.*	Includes at least **one** main character with non-specific labels only. ***Note**: Only code each character one time.* **Examples** *a) "Once there was <u>a boy</u>."* *b) "<u>The boy</u> was walking."*	Includes **one** main character **with a specific name** for the character. **Examples** *a) "Once there was a boy <u>named Charles</u>."*	Includes **more than one** main character **with specific names**. **Examples** *a) "Once there was a boy named Charles and a girl <u>named Mary</u>."*
Setting *A setting is any reference to a place or time in a narrative.*	No reference to a specific or general place. **Examples** ***a)** "The boy and the girl were walking."*	Includes reference to a general place **or** time. **Examples** *a) "The boy and the girl were <u>outside</u>.* *b) It was <u>daytime</u>.* *c) One <u>day</u>, they went to the park*	One or more references to specific places **or** times. **Examples** *a) "Once there was a boy and a girl walking in <u>central park</u>."* *b) They were walking at 10:00 at night.*	

Narrative Element	0 POINTS	1 POINT	2 POINTS	3 POINTS
Initiating Event *An initiating event is any reference to an event or problem that elicits a response from the character(s) in a narrative.*	An event or problem likely to elicit a response from a character **is not stated**. **Examples** *a) The girl looked at the boy. The boy and girl were walking in the park.*	Includes at least **one** stated event or problem that is likely to elicit a response from a character, **but there is no response directly related to that event**. **Examples** ***a)*** *"The girl was walking in a park and saw a spaceship land* ***(event/problem)*** *and she saw some aliens, and she saw a dog, and a table and...."*	Includes at least **one** stated event or problem **that elicits a response from the character(s)**. **Examples** ***a)*** *"The girl was walking in a park and saw a spaceship land and she saw some aliens* ***(IE)****. The girl started to run away* ***(Action)****."*	**Two or more** distinct stated events or problems **that elicit a response from the character(s)**. **Examples** *a) "The girl was walking in a park and saw a spaceship land and she saw some aliens* ***(IE- 1)****. The girl started to run away* ***(Action)****. But while she was running,* her shoe got stuck in a hole ***(IE-2)****. She quickly knelt down and took off her shoe to get unstuck* ***(Action)****."*

Internal Response *An internal response is any reference to information about a character's psychological state including emotions, desires, feelings, or thoughts.*	**No overt statement** about a character's psychological state.	**One overt statement** about a character's psychological state **not causally related to an event or problem**. **Examples** ***a)*** *"The dog was sad, the girl was happy."*	**One or more overt statements** about a character's psychological state **causally related to an event or problem**. **Example** *a) "The aliens' landed. Sara saw the ship and was terrified."*	

Narrative Element	0 POINTS	1 POINT	2 POINTS	3 POINTS
Plan *A plan is any cognitive verb reference that is intended to act on or solving an initiating event.* *It must include a "cognitive verb" that indicates a plan.* ***Note:** The plan and the Action/Attempt can share the same clause (see 2 POINTS example b)*	**No overt statement** is provided about the character's plan to act on or solve the event or problem. **Examples** *a) The girl was very excited and she ran out to meet the aliens.*	**One overt statement** about how the character might solve the complication or problem. **Examples** ***a)** "The girl thought that it would be neat to go and meet the aliens."*	**Two overt statements** about how the character might act on or solve the event(s) or problem(s). **Examples** *a) "The girl was very excited and she told the boy that she wanted to go meet the aliens."* *b)"The boy was very scared so he decided to sneak away quietly."*	**Three or more overt statements** about how the character might act on or solve the event(s) or problem(s).
Action *Actions are taken by the main characters but are not directly related to the IE.* *Attempts are taken by the main character(s) that are directly related to the IE.*	No actions are taken by the main character(s), **Examples** *a) There is a girl. There is a boy. It is sunny.*	Actions by main character **are not** directly related to the IE. **Examples** *a) "The boy and the girl were walking in a park.* *b) "They saw a boy alien waving."*	Attempts by main character **are** directly related to the IE. **Examples** *a)"The girl thought that it would be neat to go and meet the aliens so she got away from the boy and walked out on the grass.*	

Narrative Element	0 POINTS	1 POINT	2 POINTS	3 POINTS
Complication *A complication is an event that prohibits the execution of a plan or action taken in response to an initiating event.* **Note:** *A complication can also be a second initiating event. In this case code both a complication and initiating event.*	**No** complications.	**One** complication that prohibits a plan or action from being accomplished. **Example** *a) The spaceship landed. The girl decided to get away from the aliens and started running from the spaceship. While she was running, her shoe got stuck in a hole . She could not get away from the aliens."*	**Two** distinct complications that prohibit plans or actions from being accomplished. **Examples** *"The girl was walking in a park and saw a spaceship land and she saw some aliens (**IE- 1**). The girl started to run away (**Action-1**). But while she was running, her shoe got stuck in a hole (**Complication-1 / IE-2**). She quickly knelt down and took off her shoe to get unstuck (**Action-2**) but she was shaking too much to get her shoe off (**Complication - 2**)."*	

Narrative Element	0 POINTS	1 POINT	2 POINTS	3 POINTS
Consequence *A consequence resolves the problem or does not resolve the problem. It must be related to the IE and explicitly stated.* ***Note:** A consequence for one episode can often be the IE for another.*	**No** consequence to the action/attempt is explicitly stated. **Example** *a) "She got away from the boy and walked out onto the grass." The alien girl had a dress on.*	**One** consequence **Example** *a) The spaceship landed. The aliens were happy to see her and cried when they flew away.*	**Two** consequences **Examples** *a) They told their parents the spaceship was in the park. But their parents didn't believe them. When they took their parents to the park the spaceship was gone.* *b) The boy wanted a frog. He went to the woods to find one. <u>He couldn't find a frog.</u>*	**Three or more** consequences

Formulaic Markers *A formulaic marker is any standard utterance used to mark the beginning or ending of a narrative.* ***e.g.,** The end, once, once upon a time, they lived happily ever after etc.*	**No** formulaic utterances	**One** formulaic utterance **Example** *a) Once upon a time*	**Two or more** formulaic utterances **Example** *a) Once upon a time…The end.*	

Narrative Element	0 POINTS	*1 POINT*	*2 POINTS*	3 POINTS
Temporal Markers ***e.g.,*** *when, next, then, immediately, instantly, after, again, already, always, before, lately, now, once, presently, rarely, today, weekly, while*	**No** temporal Markers	**One** temporal marker **Examples** *a) The girl walked over to the aliens. Then they all ate some lunch.* *b) After the aliens landed, the girl screamed.*	**Two or more** temporal markers **Example** *a) When the girl saw the aliens, she ran out to meet them. She already knew they would be nice.*	
Causal adverbial clauses *e.g., because, since, so that, therefore, as a result, consequently, thus, hence etc.* *Note: causal adverbs do not have to occur in concurrent sentences*	**No** causal adverbial clauses	**One** causal adverbial clause **Example** *a) The aliens were not nice to the girl because they were scared.*	**Two or more** causal adverbial clauses **Example** *a) The aliens were not nice to the girl because they were scared. Since they were mean, she ran away.*	

Narrative Element	0 POINTS	*1 POINT*	*2 POINTS*	3 POINTS
Knowledge of dialogue *Knowledge of dialogue is registered by a comment or statement made by a character or by characters engaging in conversation.*	**No** dialogue	**One** character makes a comment or statement **Examples** a) *He said "Ow"* b) *He said "Don't come over here/"*	**Two or more** characters engage in conversation **Example** *a) He said "Oh look, there is an alien" and she said "Oh, let's go to see them.."*	
Narrator Evaluations *Narrator evaluations are any explanation provided in the story to justify why an action or event took place.* *e.g. because, since, so, and in order to.*	**No** narrator evaluations	**One** narrator evaluation **Example** *a) She ran up to say hello to the alien <u>because she always wanted to meet one.</u>*	**Two or more** narrator evaluations **Examples** *a) She knew that it was an alien spaceship <u>because everyone knows about UFOs.</u>* *b) He wanted to run from the aliens <u>since they were his worst nightmare.</u>*	

Appendix 10 Invitation letter to school principal

UNIVERSITY OF OXFORD

DEPARTMENT OF EDUCATION

15 Norham Gardens, Oxford OX2 6PY
Tel: +44(0)1865 274024 Fax: +44(0)1865 274027
general.enquiries@education.ox.ac.uk www.education.ox.ac.uk

XXX Primary School

Address

Dear Principal,

I am writing to enquire about conducting research in your [primary school name]. This study is part of my doctoral programme at the Applied Linguistics Group, Department of Education, University of Oxford, and is entitled *The assessment of Chinese EFL young learners' oral and aural language skills at primary school level.* I am a Ph.D. student at the University of Oxford, Department of Education, under the supervision of Professor Victoria Murphy and Dr. Maria Evangelou. This project has been reviewed by and received ethics clearance through the University of Oxford Central University Research Ethnics Committee.

I am interested in examining the English language outcomes of Chinese young EFL learners at primary school level. To carry out this research project, I would like to invite two classes from Grade 4 and Grade 6 in your school to participate in [the confirmed research date]. The classroom observation will be conducted for two weeks. With teachers' permission, I will take video recordings of teachers' instruction at classroom and these recordings will be erased after the thesis is completed. The assessment will be scheduled to ensure minimal disruption to your regular school activities (2-3 week maximum). The selected measures will assess various English language skills (listening, speaking, reading and writing). There will be individual speaking task and a class questionnaire on attitude to learning. These tasks are all very short (approx. 5-10 minutes each) and child-friendly.

The participation of both teachers and students is entirely voluntary. Students' parents/guardians will be asked to give written consent on their behalf. All collected data will be kept confidential, and all raw data will be destroyed when the study is completed. Any reports about the project will contain no reference that could identify the school, teachers, parents or students who have been involved.

I hope that you will agree to take part, and if you have any queries or concerns, please do not hesitate to contact me via email (xi.lian@education.ox.ac.uk) or phone (+44 (0)7411721270/+86 (0)15978622329) or my supervisors: Professor Victoria Murphy (victoria.murphy@education.ox.ac.uk) and Dr. Maria Evangelou (maria.evangelou@eduaction.ox.ac.uk).

I am happy to meet you in person to discuss the arraignments in detail at your earliest convenience. Thank you very much, and I look forward to meeting you soon.

Yours sincerely,

Xi Lian,
Dphil Candidate, Department of Education
University of Oxford, Email: xi.lian@education.ox.ac.uk
Tel: +44 (0)7411721270/ +86 (0)15978622329
W: http://www.education.ox.ac.uk/about-us/directory/xi-lian/

Appendix 11 Information sheet for teachers

UNIVERSITY OF OXFORD

DEPARTMENT OF EDUCATION

15 Norham Gardens, Oxford OX2 6PY
Tel: +44(0)1865 274024 Fax: +44(0)1865 274027
general.enquiries@education.ox.ac.uk www.education.ox.ac.uk

The assessment of Chinese EFL young learners' oral and aural language skills at primary school level

Information for teachers

This study is part of my doctoral programme at the Applied Linguistics Group, Department of Education, University of Oxford, under the supervision of Professor Victoria Murphy and Dr. Maria Evangelou.

Invitation

You are invited to take part in a research study. Before you decide to participate, it is important to understand why the research is being conducted and what your participation entails. Please take time to read the following information carefully. Please ask if there are any aspects of the project that are unclear or if you would like more information. Take time to decide whether or not you would like take part in this research.

What is the project about?

This study aims to investigate the Chinese EFL young learners' oral and aural language skills at Level 1 and Level 2 after the first few years' English learning at school. Specifically, this study will look into how teachers actually conduct their teaching of English listening and speaking according to the national curriculum; what classroom activities and tasks are they using; what kind of assessments do teachers usually use for students in Grade 4 and Grade 6.In addition, young children's feeling (e.g. learning attitudes and motivation) toward learning English and its relationship with learning outcomes will be also explored.

What will be done and what will I be asked to do in the study?

This study will be carried out in [the confirmed research date], you will be invited to answer teacher questionnaire and your class will be observed. The classroom observation will be last two weeks, with your permission, I will video-record your instruction and may take notes when necessary to document what and how you conduct you teaching and assessment.

Are there any potential risks in taking part?

This project has been reviewed by, and received ethical clearance through, the University of Oxford Central University Research Ethics Committee. If you decide to take part in the research you will be asked to sign a consent form to show that you have agreed to participate. There are no known or foreseen risks to participant in this study. Your participation is entirely voluntary. You may withdraw from the study without reasons at any point, and you may request to remove all or part of your data. You are not obliged to answer any question that you find objectionable or that makes you feel uncomfortable. Your name will be replaced by the pseudonym and no identifying information will be included in the data file and thesis.

What are the benefits in taking part?

Your participation, as part of this study, will help researchers and school teachers better understand the teaching and assessment in listening and speaking at primary level in China.

Will the research be published?

The research will be written up as a thesis. On successful submission of the thesis, it will be deposited both in print and online in the University archives, to facilitate its use in future research.

Confidentiality

The information collected will be treated confidentially. Reports will not identify individuals or organisations by name, but talk about what was found more generally. The data will be kept securely and destroyed after the research is completed.

Xi Lian
Dphil Candidate,
Department of Education; University of Oxford,
Email: xi.lian@education.ox.ac.uk ;
Tel: +44 (0)7411721270/ +86 (0)15978622329
W: http://www.education.ox.ac.uk/about-us/directory/xi-lian/

Appendix 12 Information sheet for parents/guardians

UNIVERSITY OF OXFORD

DEPARTMENT OF EDUCATION

15 Norham Gardens, Oxford OX2 6PY
Tel: +44(0)1865 274024 Fax: +44(0)1865 274027
general.enquiries@education.ox.ac.uk www.education.ox.ac.uk

The assessment of Chinese EFL young learners' oral and aural language skills at primary school level

Information for parents/guardians

This study is part of my doctoral programme at the Applied Linguistics Group, Department of Education, University of Oxford, under the supervision of Professor Victoria Murphy and Dr. Maria Evangelou.

Invitation

Your child is being invited to take part in a research study. Before you decide to allow your child to participate, it is important to understand why the research is being conducted and what your child's participation entails. Please take time to read the following information carefully. Please ask if there are any aspects of the project that are unclear or if you would like more information. Take time to decide whether or not you would like your child to take part in this research.

What is the project about?

This study aims to investigate the Chinese EFL young learners' oral and aural language ability at Level 1 and Level 2 after the first few years' English learning at school. In addition, young children's feeling (e.g. learning attitudes and motivation) toward learning English and its relationship with learning outcomes will be also explored.

Why has my child been chosen?

For this study, we are seeking for a group of Chinese primary school students who are learning English as a foreign language. Your child belongs to this group and has simply been chose for that reason. His/her classmates in the class are also invited to participate in the study.

What will be done and what will my child be asked to do in the study?

This study will be carried out in [the confirmed research date], your child's English achievements will be measured by a set of classroom tasks (listening, reading and writing) and your child may also be invited to attend the individual speaking task that would be conducted by the researcher at the school. This speaking task will be audio recorded and later transcribed. In addition, your child will also be invited to complete a simple questionnaire on his/her attitude toward English language learning.

Dose my child have to take part? Are there any potential risks in taking part?

This project has been reviewed by, and received ethical clearance through, the University of Oxford Central University Research Ethics Committee. If you decide your child can take part in the research you will be asked to sign a consent form to show that you have agreed to your child's participation. Even if you do allow your child to participate, he or she will be allowed to withdraw at any time if they wish.

Your child does not need to answer questions that you do not want him/her to. Your child will also be, prior to the study, informed orally of the aims and procedures of the research and asked orally whether he/she is willing to take part. His/her participation will be completely anonymous, and every effort will be made to preserve confidentiality. The results of the tasks will not affect his/her academic results in any way.

What are the benefits in taking part?

Your child's participation, as part of this study, will help researchers and school teachers better understand the learning achievements, learning attitude and motivation of the Chinese young EFL learners at primary school level.

Will the research be published?

The research will be written up as a thesis. On successful submission of the thesis, it will be deposited both in print and online in the University archives, to facilitate its use in future research.

Confidentiality

The information collected will be treated confidentially. Reports will not identify individuals or organisations by name but talk about what was found more generally. The data will be kept securely and destroyed after the research is completed.

Xi Lian, Dphil Candidate,
Department of Education,
University of Oxford,
Email: xi.lian@education.ox.ac.uk;
Tel: +44 (0)7411721270/ +86 (0)15978622329

Appendix 13 Consent form for teachers



UNIVERSITY OF OXFORD

DEPARTMENT OF EDUCATION

15 Norham Gardens, Oxford OX2 6PY
Tel: +44(0)1865 274024 Fax: +44(0)1865 274027
general.enquiries@education.ox.ac.uk www.education.ox.ac.uk

The assessment of Chinese EFL young learners' oral and aural language skills at primary school level

Consent form for teachers

This study aims to examine the Chinese EFL young learners' oral and aural language skills at Level 1 and Level 2 after the first few years' English learning at school. This is a study undertaken by Xi Lian, a Ph.D. student in the Applied Linguistics Group, Department of Education, University of Oxford.

Your school has agreed to take part in a study run by Oxford University looking at teaching and assessment practice in English listening and speaking at primary school level.

- If you agree to takes part, a researcher would come and visit your class at school, and you will be asked to answer teacher questionnaire and your lesson will be observed.
- If you are happy to take part, please fill in the form below and return to the researcher.
- To find out more about the project, please read the attached information sheet. If you have any queries or concerns, please do not hesitate to contact me via email (xi.lian@education.ox.ac.uk) or phone (+44 (0)7411721270/ +86 (0)15978622329) or my supervisors, Professor Victoria Murphy at victoria.murphy@education.ox.ac.uk, and Dr. Maria Evangelou at maria.evangelou@eduaction.ox.ac.uk.

- - - -

Appendix 14 Consent form for parents/guardians

UNIVERSITY OF OXFORD

DEPARTMENT OF EDUCATION

15 Norham Gardens, Oxford OX2 6PY

Tel: +44(0)1865 274024 Fax: +44(0)1865 274027

general.enquiries@education.ox.ac.uk www.education.ox.ac.uk

The assessment of Chinese EFL young learners' oral and aural language ability at primary school level

Consent form for parents/guardians

This study aims to examine the Chinese EFL young learners' oral and aural language ability at Level 1 and Level 2 after the first few years' English learning at school. This is a study undertaken by Xi Lian, a Ph.D. student in the Applied Linguistics Group, Department of Education, University of Oxford.

Your child's school has agreed to take part in a study run by Oxford University looking at the learning outcomes of Chinese EFL young learners at primary school level.

- If your child takes part, a researcher would come and visit them at school, and do some language tasks with them.
- **If you are happy for your child to take part, please fill in the form below and return it to your child's class teacher as soon as possible.**
- To find out more about the project, please read the attached information sheet. If you have any queries or concerns, please do not hesitate to contact me via email (xi.lian@education.ox.ac.uk) or phone (+44 (0)7411721270/ +86 (0)15978622329) or my supervisors, Professor Victoria Murphy at victoria.murphy@education.ox.ac.uk, and Dr. Maria Evangelou at maria.evangelou@eduaction.ox.ac.uk.

- - - -