The Preparation of Assessment Literate Teachers: Examining Language Testing and Assessment Courses in China and New Zealand

Wenjing Yao

A thesis submitted in complete fulfilment of the requirements for the degree of [Doctor of Philosophy in Education], The University of Auckland, 2020.

Abstract

This thesis investigates the language testing and assessment courses (LTACs) in different cultural and academic contexts (Chinese undergraduate, Chinese postgraduate and New Zealand postgraduate) with an intention to: identify similarities and differences in terms of course construction and implementation; examine how LTACs develop students' assessment literacy in China; and provide implications in contrasting contexts for assessment literacy preparation in teacher education programmes.

An exploratory mixed-method research comprising two sequential phases of data collection and three inter-connected studies was designed and conducted. Study 1 employed content analysis to 20 LTAC course syllabi and identified different foci across the three contexts: Chinese undergraduate LTACs put a heavy weighting on language testing; Chinese postgraduate LTACs incorporated the teaching of formative and classroom assessment in addition to testing; and New Zealand (NZ) postgraduate LTACs focused on assessment as an umbrella term covering testing and other forms of assessment. Study 2 applied thematic analysis to interviews collected from the 20 course instructors and revealed that: 1) Chinese and NZ instructors all adopted a practice-based approach in teaching and were all confronted with challenges from time limits and students' limitations; 2) Chinese and NZ instructors held contrasting attitudes towards policy and curriculum guidance; and 3) instructors' conceptions of assessment and definitions of assessment literacy varied due to external contextual factors and individual differences, causing differences and similarities in their course design. Study 3 surveyed senior pre-service teachers from the Chinese undergraduate context and statistically analysed the impact of LTACs on their conceptual model of assessment literacy which includes three components: conceptions of assessment, self-efficacy of assessment, and practices of assessment. The survey results showed that: 1) Self-efficacy functioned as a mediator between conceptions and practices of assessment; 2) only one third of participanting pre-service teachers had experienced an LTAC and LTAC experience made little difference to pre-service teachers’ conceptual model of assessment literacy.

Collectively, the findings informed a model of contextual factors that influence course construction and assessment literacy preparation that are applicable to different contexts. This thesis supports previous research that assessment literacy preparation is a social process in which many agents are involved and is influenced by the external educational environment. Additionally, the thesis shows that the power of external environments is mediated by differences in instructors’ individual experiences and backgrounds.

The findings, however, also showed, at least in China, the inadequacy of one single assessment preparation course for assessment literacy development. Based on these findings and discussions, theoretical and practical implications were raised around assessment preparation courses, teacher education programmes, and the implementation of educational policies.

Acknowledgements

There are a number of people and parties to whom I would like to extend my warmest and most sincere thanks for making the completion of my PhD study possible. My first thanks go to the China Scholarship Council who funded my 4-year doctoral study; and then to my dear supervisors Professor Gavin Brown and Associate Professor Mary Hill who unwaveringly supported me, mentored me and pulled me back on track from time to time along the PhD journey. They taught not only by supervision meetings but more by the example of how to be a meticulous yet caring scholar and educator.

A million thanks also go to the LTAC instructors who participated in the study in China and New Zealand. Some of them are well-known field experts whose interviews are precious lessons for a novice teacher-researcher like me, while others are my peers who generously shared their pearls of wisdom. Their participation in this work counts not only as a vital and indispensable part of this PhD project, but also a life-long memory that humbled me, inspired me with the "keep-on-learning" attitude.

Hearty thanks go to my bosom friends within and beyond the PhD circle: my H302 office mates, the LDPP luncheon group, the MAY ladies, Dr. Graeme Sterne, Dr. Robyn Barry, Dr. Scott Ray, , Dr. Libby Limbrick, and Dr. Bing Eric Mei.

Last but not least, I would like to say "谢谢 (*xiexie)*" to my mom, my in-law parents, who, alternately, came all the way from the other hemisphere to help me with child-minding and housework so that I could focus on study. To my husband Peng, who supported me by taking all the household duties and encouraging me from time to time, and also to Amy, our daughter, my biggest drive to excel. This piece of work is the best example of 'mummy's never-give-up spirit' that I wish to pass on to you.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

The research topics and questions for this thesis stem from my personal curiosity and inquiry of language assessment based on my teaching and learning experiences. As an ESOL (English for speakers of other languages) teacher who was born, raised and educated in China but received doctoral-level academic and research training in New Zealand (with research area in language assessment and teacher education), I realized that my conceptions, self-efficacy, and practices of assessment have changed a great deal since I came to New Zealand.

China and New Zealand (NZ for short), located in opposite hemispheres, vary across a wide range of dimensions, one of which is the assessment culture. China is quite intense in regard to assessment while NZ is less highly-structured. During my schooling in China, I was tested very frequently and spent most of time preparing for the high-stakes entrance exams throughout; I did not have the chance to receive a language testing and assessment course (hereafter LTACs) while studying for my bachelor's and master's degrees in two universities in China. Consequently, my conceptions of language assessment was narrowed to language testing. During my doctoral study in NZ, I learned and experienced multiple forms of assessment; studied the cutting-edge literature on educational/language assessment and was assessed from time to time with feedback from my supervisors throughout my study. I now view assessment as a comprehensive and interactive system that's more than what testing can explain.

Thus, this thesis integrated my research interests of language assessment, assessment literacy preparation, and teacher education and set out to explore how ESOL teachers were prepared in terms of assessment literacy in these contrasting contexts with a focus on the language testing and assessment courses. I believed it would be stimulating and meaningful to see how current ESOL teachers have been prepared in assessment, which in turn may predict the future of teachers' assessment practices in China and NZ. In the following sections in this chapter, the research background, the significance of the research, and the thesis structure are presented.

1.1 The Background

With the on-going advancement of globalization and the rising position of English as lingua franca, both China and New Zealand have seen a huge surge in the demand for English language learning as either a foreign language or a second language (EFL or ESL)[1]. In China, English as a second language is explicitly included as a compulsory subject from year 3 in primary school by the Ministry of Education (Chinese Ministry of Education, 2011). It is also implicitly regarded as a stepping-stone to better domestic and international educational opportunities and greater job prospects (Hu, 2003). Similarly, New Zealand has also seen a dramatic increase in the number of international students and immigrants eager to learn English well in order to live and study in this country since the last decade (Campbell & Li, 2008; Li, 2004). Among the international students studying in NZ, Chinese students account for a significant portion. According to Education New Zealand (2018), there are 106,021 international students enrolled in school sectors as at 31 August, 2017 with 33% of the total enrolment coming from China.

There are huge differences between the two nations in terms of ESOL education. To start with, ESOL classroom teaching in China is largely teacher-centered with grammar-translation as the main method (Miao, 2007; Xu & Connelly, 2009) whereas in New Zealand the ESOL teaching focuses more on the intercultural communicative approach that requires learners' active participation (Campbell & Li, 2008; Newton, 2009). In addition, the assessment culture in the educational system between China and New Zealand is also

[1] The major difference between EFL and ESL lies in the use of English in different contexts: the former refers to teaching English in a non-English speaking country where English is learned as a foreign language (e.g., China), while the latter means teaching English to non-native English speakers in an English-speaking country where English is spoken as the second language of the learner (e.g., New Zealand). As recommended by Nation (2012) this thesis uses the term ESOL to refer to both for simplicity and convenience.

distinctive. China, deemed as the representative of Confucius-heritage country, is known for its test-oriented culture (Hu, 2005; Liao, 2004) in which good educational resources are fiercely sought after through intensely competitive high-risk examinations, serving as a guarantee for future success (Chen & Brown, 2013, 2016; Gu, 2014; Jin, 2010). The purpose of assessment in China is still largely for selection and competition, thus making assessment in education highly centralized and examination focused with high stakes (Cheng, 2008). Comparatively New Zealand features a pro-formative assessment educational environment with less focus on tests and exams (Brown, 2011). There is no compulsory, large-scale, high-stakes examinations in primary and intermediate education, but a standards-referenced national exam system (National Certificate of Educational Achievement) for entrance to tertiary education where only a portion of assessment involves examinations (Crooks, 2002; Philips, 2000). Assessment in New Zealand education is espoused as being for improvement and adjustment (Crooks, 2004; Crooks, 2010), with lower stakes and being less centralized than in China. High stakes testing in New Zealand, while present, is less dominant than in China.

Despite the differences between the two countries, the similar large demand for English language learning resulted in the consequent demand for ESOL teachers in both countries. Yet, contrary to the large demands, the quality of ESOL teachers needs improvement, especially in terms of their knowledge and skills of assessment, referred to as assessment literacy (Berry & O'Sullivan, 2016). Teachers' assessment literacy plays a decisive role in their assessment practices and is vital to teachers (Boyles, 2006; Taylor, 2009). According to Stiggins (2002), teachers with adequate high-level assessment literacy make teaching and learning more effective and engaging. Researchers have worked hard to define, classify and measure assessment literacy (e.g., Deluca & Klinger, 2010; Mertler, 2004; Mertler, 2005; Popham, 2004; Stiggins, 1995; Xu & Brown, 2016), however, there is little research with specific focus on assessment preparation courses (the LTACs) in ESOL teacher education programmes and even scarcer study with a comparative view in disparate cultural and academic settings.

This thesis addresses this research gap by investigating the LTACs in ESOL teacher education programmes in China and New Zealand across three specific contexts: Chinese undergraduate, Chinese postgraduate and NZ postgraduate. The Chinese undergraduate courses are part of the four-year Bachelor of TESOL (teaching English for speakers of other languages) programmes. The Chinese postgraduate courses are embedded in the Master of TESOL programmes, whereas the NZ postgraduate LTACs are part of either a Master of TESOL or Postgraduate Diploma in TESOL programmes, which all last for two years full time.

The research aims to: 1) enhance the understanding of course construction and implementation in different contexts; 2) examine the impact of assessment preparation courses on student teachers' literacy development; and 3) provide theoretically and empirically based recommendations for various stakeholders involved.

1.2 The Significance

Researchers and educators worldwide support the notion that effective assessment informs teachers' teaching and improves students' learning (Black & Wiliam, 1998; Harlen & James, 1997; Sadler, 1998). Assessment literacy preparation in teacher education programmes is theoretically and practically essential because: 1) Teachers with a solid foundation in language assessment are more prepared to integrate assessment with instruction effectively and efficiently (Brindley, 2001; Inbar-Lourie, 2008); and 2) assessment is a widespread aspect of most educational systems in the world today, without which valid and effective teaching and learning is not possible (Biggs, 2003; Eyers, 2014). As language teachers spend a great amount of time in testing and assessment, it is important that they should be equipped with ample knowledge of the field before stepping into their professional teaching career (Stiggins, 2002). Nevertheless, in contrast to the exponential increase of English learners and the attention given to ESOL teacher education worldwide (Hu, 2005; Wu, 2001), inadequate attention has been paid to ESOL teachers' assessment literacy development (Fulcher, 2012; Lam, 2015). As Berger (2012) states, "given the importance of language assessment in today's world, it is surprising if not alarming how little is known about it and how significant a role language assessment plays in teacher education programmes" (p. 57).

A teachers' assessment literacy is a combined result of his or her prior learning experience as a student, teacher preparation and professional development, along with the teaching environment s/he is in (Brown, et al., 2009). As the starting point for a teacher's professional journey, teacher education paves the way for assessment literacy development both theoretically and practically (Brindley, 2001; DeLuca & Klinger, 2010; Deluca et al., 2013; Deluca & Lam, 2014; Mertler, 2004). Within these programmes, assessment courses play a pivotal role. Undertaking a language assessment course within teacher education is a first and fundamental step of equipping ESOL teachers with adequate assessment knowledge. This lays the foundation for their future assessment practices in the real classroom as well as for future professional development (Hill et al., 2010). With a solid background in language assessment received at teacher education programmes, teachers could be better positioned to use assessment to improve teaching and learning. However, despite the importance of preparing language teachers to be assessment capable, "the process towards an assessment-literate educational culture has been slow and inadequate" (Coombe, O'Sullivan, & Stoynoff, 2012, p.21).

Another significant aspect of this research lies in the comparison of LTACs in different cultural and academic contexts. This is supported by the following two reasons: understanding the social dimensions of assessment in language education and shedding light on each comparative site. To start with, language assessment has a social dimension, or is socially oriented (McNamara, 2001; Roever & McNamara, 2006). The investigation into the interaction between social-cultural, educational context and assessment literacy preparation in LTACs help us understand the social dimensions of assessment. Boaler and Humphreys (2005) emphasised the importance of viewing language assessment as a social activity that entails a reciprocal relationship between language assessment and its social influence: Societal disparities exert an influence not only on the operation of language assessment, but also on the shaping of conceptions and beliefs of student teachers during their education programmes (Xu & Brown, 2016). Moreover, the impact of language assessment has widened. The assessment results are used for measuring learners' ability to use language in social settings. In addition to informing educational outcomes, language assessment is also used for selection in employment and immigration contexts (McNamara, 2001). The development of teachers' assessment literacy is an embedded activity shaped by the internal local context with teachers, students and the community who are recognized as meaningful assessment co- operators (Leung, 2004; Lynch, 2001).

The comparative study can also provide us with theoretical and empirical understanding and implications for each party involved. As argued by Bray, Adamson, and Mason (2014), "Many people who undertake comparative study of education find not only that they learn more about other cultures and societies but also they learn more about their own" (p. 35). The benefits of comparative study depend largely on "how the results of the research are used with the power to action them" (p. 36). As suggested by Porter and Gamoran (2002), one of the most frequently quoted benefits of international comparative education research is that, "education in one country can be better understood in comparison to education in other countries" (Porter and Gamoran, 2002, p. 7).

To conclude, assessment literacy is an important and indispensable aspect of teacher education, as well as a social process in which many agents are involved. Investigating LTACs in comparative contexts enables a better understanding and explanation of the social dimension of language assessment, the nature of teacher assessment literacy preparation and its development within teacher education programmes.

1.3 The Thesis Structure

This thesis is in seven chapters. Chapter 1 provides the research background, explains its significance and introduces the thesis structure. The major significance lies in the importance of assessment literacy preparation and the comparative nature of the study.

Chapter 2 synthesizes and summarizes previous significant studies related to the research and presents the theoretical framework for analysis. Key terminology used for this thesis is defined; educational and assessment environments in China and New Zealand are illustrated; and key issues in assessment literacy preparation in teacher education programmes are discussed. Based on the literature review and the research interests, research questions are then raised at the end of the chapter.

Chapter 3 describes the methodology employed in this thesis including its research paradigm, research design, data collection and data analysis methods. An exploratory sequential mixed-methods research composed of two phases of data collection and three inter-linked studies was designed and implemented to address the research questions. Phase-I collected qualitative data from 20 course syllabi and 20 course instructors' interviews, while Phase-II collected quantitative data on Chinese undergraduate student teachers' conceptions, self-efficacy and practices of assessment via web-based on-line survey. The three inter-linked studies form a research triangulation that contributes to the overall understanding on how LTACs are developed and implemented under different contexts and how they helped shape student teachers' assessment literacy.

Chapter 4 reports the findings of Study 1: a content analysis of 20 course syllabi. Study 1 serves as a starter and explores the essential course components, namely, course objectives, contents and assessment plans as written in the syllabi to identify the focus of the courses and assessment literacy preparation.

Chapter 5 reports the findings of Study 2: a thematic analysis of 20 course instructors' interviews. Study 2 is a follow-up of Study 1 and provided further information on LTACs construction and implementation as seen from the instructors' perspectives. The thematic analysis revealed four themes regarding the similarities and differences in course construction and implementation.

Chapter 6 reports the findings of Study 3: a quantitative analysis (the major statistical analytic tools are confirmatory factor analysis and structural equation modelling) of Chinese undergraduate student teachers' survey. Based on the findings of Study 1 and Study 2, Study 3 explored the structural relationship between the three components of assessment literacy and examined the impact of LTACs on students' conceptual model of assessment literacy.

Chapter 7 integrates findings from the three studies and discusses the major issues concerned. A model of the contextual factors that influence the assessment preparation courses and students' assessment literacy development is proposed. This chapter also explains the theoretical and practical implications and contributions of this thesis, together with limitations and recommendations for future study.

Chapter 2 Literature Review

"As a pre-requisite to an empirical study, the literature review identifies gaps in knowledge and justifies your forthcoming research study" (Aveyard, 2014, p. ix). This chapter provides a review of pertinent literature to provide a theoretical and practical background to the empirical research in this thesis. It starts by defining and discussing key terminology, and proceeds by introducing the external assessment environments in China and NZ as research background. In the following section, prior studies of assessment literacy in teacher education programmes are synthesized and summarized based on the following three strands: 1) a discussion of definitions and frameworks of assessment literacy in teacher education; 2) assessment preparation in teacher education programmes and 3) the measurements used to evaluate teacher assessment literacy. This chapter concludes with the principal and subsidiary research questions for this thesis.

In searching for relevant and significant literature pertaining to the research, I searched major academic databases including *EBSCO, ERIC, Google Scholar, ProQuest, Dissertation and Theses database* to identify articles, books and other resources with key words including, but not limited to, "educational assessment, language assessment, assessment literacy/competency/ability/capability, assessment reparation/training/development, teacher conceptions and practices of assessment". In addition, important and frequently cited references from journal articles, books, websites and doctoral thesis were further searched and studied. Alerts of the latest publications on language assessment and teacher education from *Google Scholar* were set to be sent directly to the researcher's email to keep up with the research trends. A general principle for selection is that literature works are mostly published after the 1990s to keep up with the latest developments in the fast-changing research field of education and assessment (Rowley & Slack, 2004).

2.1 A Working Definition of Key Terms

It is important to use words with precision in developing arguments relating to topics as complex as assessment (Damon, 2007; Wiliam, 2011), thus, this section defines and clarifies the fundamental terms that appear throughout this thesis to make sure the readers and the researcher have a shared understanding of the meaning of terms. To address this issue, definitions of assessment in general education as well as in language education are explained, summative and formative assessment are compared and frequently used concepts in this thesis are defined for this thesis.

2.1.1 Educational assessment and language assessment.

Assessment has become a focal point in both general education and language education in the last couple of decades (e.g., Kumaravadivelu, 2012; Popham, 2011; Rea-Dickins, 2007). Sharing the fundamental concepts and principles with educational assessment, language assessment has its roots in the legacy of educational assessment (Bailey & Curtis, 2015; Freeman & Johnson, 1998) while its specialty is determined by its disciplinary characteristics, such as its communicative nature and social-cultural dimension (Johnson, 2006; McNamara, 2001).

2.1.1.1 Educational assessment.

The word “assessment” first came into use in the field of education after the Second World War (Nelson & Dawson, 2014). There are numerous definitions of educational assessment used by key researchers. Some emphasize it as a process of “…gathering, interpreting, recording and using information about students’ responses to educational tasks” (Lambert & Lines, 2013, p.7), while others define it as a method “…that (is) used to determine what students know and are able to do before, during and after instruction” (Green & Johnson, 2010, p.14). A comprehensive definition of educational assessment regards it as both a method and a process. As Brown (2018) stated in his work: “Educational assessment refers to the set of methods and processes by which evidence about student learning is designed, collected, scored, analysed, and interpreted. These processes are meant to support decisions about teaching, learning, administration, policymaking, and accountability (p.1).” Regardless of its form, assessment in education involves making decisions about the relevant evidence for a particular purpose, the collection and interpretation of the evidence, and communication of assessment results to the intended receivers (Harlen, 2005a). In this thesis, assessment refers to the process and methods of generating and collecting information about how, and what, students have learned for the purposes of diagnosis, evaluation, providing feedback as well as using the information for selection.

2.1.1.2 Language assessment.

Definitions of language assessment are similar to those in general education but the “given object of assessment interest is mostly the language ability” (Bachman, 2004, p. 27). According to Bailey and Brown (1998), language assessment measures the effectiveness of individual learning and facilitates teachers in tracking students’ learning trajectories and promoting the learning outcomes. Brown (2004) insisted that language assessment should be an on-going process with a set of methods to gain information about the effects of language teaching and learning. Bachman (2004)’s definition about language assessment supports these notions:

> Language assessment can be thought of broadly as the process of collecting information about a given object of interest according to procedures that are thematic and substantively grounded. A tool, or method of this process, such as a test score or a verbal description, is also referred to as an assessment (p. 7).

It should be noted that language assessment, or to be specific, assessing English as a second or foreign language differs from assessment in general education that it involves more communication of the target language and thus should be more focused on the genuine communicative function of the language (Griffiths & Parr, 2001; Savignon, 1991; Savignon, 2002). One reason may be contributed to the pedagogical characteristics of foreign language teaching, namely, the instructional practices mainly consist of dialogue, conversation or discussion. As argued by Johnson (2006), there are more frequent and direct interactions between the teachers and students than in other curriculum areas where lecturing and listening are the major classroom activity.

In addition, one of the fundamental features of language assessment is fundamentally its social nature (McNamara, 2001). Language assessment and the societal environment forms an interactive and inter-related relationship. On one hand, the impact of language assessment has been widened for purposes besides informing educational results, including for employment, for selection and for immigration (McNamara, 2001; O'Sullivan & Stoynoff, 2012). On the other hand, societal disparities exert an influence not only on the operation, administration and interpretation of language assessment, but also on the formation of language teachers' conceptions and beliefs about language assessment during their education programmes (Chen & Brown, 2016; Roever & McNamara, 2006).

Although assessment is often used interchangeably with test/testing, its conception is broader and encompasses a wider domain. Brown (2004) distinguished the concepts of testing and assessing by arguing that a test is more a technical method measuring the test-takers' content knowledge or performance in certain subject whereas assessment is more about "an ongoing process that encompasses a much wider domain" (p.4). In this way, language assessment should be regarded as an umbrella term that includes traditional testing as well as alternatives of assessment such as classroom-based assessment activities including spontaneous teacher feedback and observation, students' self and peer assessment. Bachman (2004) also explained language assessment as the process of "developing, scoring, interpreting and improving classroom-based assessments as well as selecting, administering, interpreting and sharing results of large-scale tests developed by professional testing organizations" (p.9).

2.1.2 The dichotomy of summative assessment and formative assessment.

Summative and formative assessment are the two most frequently used concepts in assessment study. Scriven (1967) first proposed the concepts of summative and formative in relation to evaluation of educational programmes by referring to the final and overall evaluation of a programme as "summative" while the evaluation "during the construction and trying out a new curriculum" as "formative" (p. 51). Bloom, Hastings and Madaus (1971) later extended the usage to the current generally accepted meanings.

This section 1) defines the two forms of assessments in terms of the timing, format, and underpinning psychological principles; and 2) compares their advantages and disadvantages; and 3) discusses the relationship between the two. It argues that 1) there has been a paradigm shift from an emphasis on summative to formative assessment in the field of education; and 2) instead of being viewed as a pair of opposite binary, they should be interpreted as interacting and complementary to each other.

2.1.2.1 Summative assessment.

Summative assessment (SA) refers to the assessment event that occurs either after a set period or at the end of the instruction. It is conducted mostly in the form of tests, exams or projects assigned with marks or grades to summarize the learning results with focus on the outcomes (Harlen, 2007; Bennett, 2010). The most frequently applied form of SA is standardized testing administered and scored in a consistent and standardized way (Popham, 1999). Behaviorist theory, claiming that behavior and performance is molded by reinforcement, stimulus-response and repeated learning is reported to have an influence on SA (Berry, 2008). The purpose of SA is usually associated with grading, selection and accountability with an emphasis on the accuracy and consistency of evidence of meanings (Brown, 2019; Wiliam & Black, 1996).

An advantage of SA is that it can provide clear guidance and great stimulus for teachers' teaching and students' learning (Earl, 2012). Also, when designed with established validity and reliability, SA can function as an efficient and effective selection tool for large-scale selection by avoiding time-consuming and sometimes biased teacher assessment (Harlen, 2005a). Nevertheless, its critics argued that learners are inclined to be passive respondents in SA (Harlen, 2005b; Kealey, 2010). With its summative nature, SA usually measures a limited range of knowledge in which high scores can be achieved through memorization and repeated drills without the desired degree of deep thinking (Biggs, 1998). Moreover, it is can be challenging to address the constituent and dynamic teaching/learning process within SA (Leung, 2004). In addition, when SA is associated with the selection and accountability purposes of assessment with high-stakes at a national or fate-determining level (e.g., the national English examination held in China for college entrance and the IELTS test for emigration purposes), it can cause negative washback effect on teaching and learning (Jin, 2010; Sapp, 2013)..

2.1.2.2 Formative assessment.

In contrast to summative assessment, the term formative assessment (FA) refers to the assessment usually conducted during the teaching process in the form of teacher observation, self- and peer-assessment, and teacher questioning (Sadler, 1998; Black & Wiliam, 2006). FA gathers information to measure the effectiveness of curriculum and teaching so as to inform decisions on how to make improvements. Sadler (1998) first applied the concepts of FA in teaching and claimed that “formative assessment is concerned with how judgements about the quality of students’ responses (performance, pieces, or works) can be used to shape and improve the students’ competence by short-circuiting the randomness and inefficiency of trial-and-error learning” (p.80). The underlying philosophy of FA is social constructivism which emphasized that knowledge construction involves social and learner-teacher interaction (Berry, 2008). It is argued that learning is constructed by learners themselves through interactions with peer students, with teachers and is influenced by the learning environment (Biggs, 1996; Lantolf, 2000; Rushton, 2005). The purpose of FA is to inform of gaps for improvement and promote learning rather than to select and make accountability (Black & Wiliam, 1998; Sadler, 1998). Students are supposed to be actively involved in formative assessment while teachers are important facilitators to guide the learning process. Consequently, both the teachers’ and the students’ roles are essential in FA (Bennett, 2010; Wiliam, 2011).

FA happens during the teaching process which enables teachers to provide timely feedback catering to students' individual needs and characteristics (Crooks, 2004). It also helps teachers and students to identify gaps for improvement (Wiliam & Black, 1996). Furthermore, it emphasizes the learners' active participation, so it helps to boost students' motivation and self-regulation for study while releasing the pressure from summative testing. However, considering its in-the-moment and on-the-fly nature of much formative assessment practice, it may lack evidence for trustworthiness and accuracy (Brown, 2019). In addition, when FA is used in line with teaching, it can be challenging to distinguish FA from pedagogical instruction (Brown, 2019).

2.1.2.3 The shift from emphasis on SA to FA.

Summative assessment in the form of written tests with standardized answers dominated the policy and education landscape before the 1990s worldwide when the major function of assessment was to select qualified students for entry into college or better opportunities (Earl, 2012; Taras, 2005). In the last two decades, however, the prevalence of social-constructivist theories of learning and curriculum reform have led to an assessment reform and a shift of focus in educational assessment in many western contexts (Banta, 2002; Shepard, 2000). The reliance on external standardized testing as the major and only assessment tool has been reduced while increasing attention has been paid to formative assessment as an effective method to improve the quality of teaching and learning (Gipps, 2011; Shepard, 2006).

The concept and practice of formative assessment was further expanded and promoted in the UK through the efforts of The Assessment Reform Group, initiated by Black and Wiliam (1998). With their seminal review of around 250 studies relating to FA, Black and Wiliam (2009) defined formative assessment as, "the practice in a classroom is formative to the extent that evidence about student achievement is elicited, interpreted and used by teachers, learners, or their peers, to make decisions about the next steps in instruction (p.9); and they concluded that "attention to formative assessment can lead to significant learning gains (p.52)."

Ever since then, FA has gained increasing attention in both general education and language education. Similarly, in language assessment, the proliferation of FA signified a change of focus from a "technical endeavor to an on-going scheme" (Rea-Dickins, 2007, p. 515). Whereas the 'technical endeavor' focuses on creating and using well-designed tests to measure learners' language learning, the 'on-going scheme' supports the learners throughout the learning process (Cummins & Davison, 2007).

2.1.2.4 The relationship between SA and FA.

A number of researchers are concerned about the lack of clarity in the distinction between SA and FA (Black et al., 2004; Black & Wiliam, 1998; Harlen & James, 1997). Within official documents great importance and emphasis has been placed on the timing as a factor for differences. FA occurs during the instruction while SA at some end point/s.

A misconception often held by teachers is that tests are only summative in nature and this is perhaps because most testing is conducted periodically or at the end of study to measure learning outcomes (Stiggins, 2002; Wiliam & Black, 1996). According to Wiliam and Black (1996) the distinction between formative assessment and summative assessment lies in not only when they take place but also in the functions they serve. As mentioned above, SA is usually conducted in the form of tests and employed at the end of the learning process and, in most cases, for the purpose of evaluation, selection, employment and classification. In contrast, formative assessment can be used before, during and after the learning process with the purpose to provide feedback to assist students' learning; it can be either testing or another forms of assessment.

Black and Wiliam (2009) also argued that the results of summative tests can be used formatively. Using information about learning from summative test results formatively can provide students with feedback that motivates learning, even though the tests are designed for the purpose of selection and gathering evidence about academic achievement. In this way, tests can function as a direction for academic success, as a guide for students helping themselves in learning and as a guide for charting their own learning. An illustration of this is the learning and assessment software designed and established in New Zealand between 2000 and 2008, known as the Assessment Tools for Teaching and Learning (asTTle) (Archer & Brown, 2013; Hattie & Brown, 2007). The asTTle system adopts a formative and validity-focused process that allows teachers to determine the assessment focus, diagnose the assessment result and adjust their teaching practice accordingly to help students learn better. This educational resource enables both improvement and reporting responses through assessing students' academic performance. The assessment tool is a database of items and tasks designed by classroom teachers, subject specialists, and curriculum experts that allows schools and teachers to generate standardized tests aligned with national curriculum statements. Its aim is to promote students' learning by describing their performance and counteracting the negative effects of compulsory national testing.

It is suggested that instead of seeing SA and FA as opposite to each other, teachers should connect formative and summative assessment and use them in harmony in line with the overall teaching and learning goals (Houston & Thompson, 2017; Lau, 2016).

2.1.3 Testing, classroom assessment and formative assessment.

Testing is the most traditional form of language assessment that has long been used as the main method for evaluating students' language levels. Language tests are usually in a written form that requires students to answer questions such as multiple choice, cloze items, short answers and the results can be checked with standardized answers (Cheng, 2005; Gipps, 2011). Testing represents a technical approach that measures performance in a standardized way and good tests require a robust mechanism to ensure the accuracy and trustworthiness so as to make valid decision based on the tests results (Brown, 2019). Nevertheless, because of the communicative nature of language learning, traditional testing is inadequate for measuring students' language performance (Richards & Rogers, 2014). Furthermore, high-stakes testing can lead to various negative washback effects to learning (see more discussion in section 2.3.1).

To combat the shortcomings of tests and assess students' language ability in a consistent and on-going way, classroom assessment (CA) has been proposed and advocated in language teaching and learning (Brown & Abeywickrama, 2010; Richards & Renandya, 2002). CA is used to refer to "those assessments developed and used by teachers in the classroom on a day-to-day basis (Stiggins & Conklin, 1992, p.1)". McMillan (2013) states that it is a "broad and evolving conceptualization of a process that teachers and students use in collecting, evaluating and using evidence of student learning for a variety of purposes, including diagnosing, student strengths and weaknesses, monitoring student progress towards meeting desired levels of proficiency, assigning grades, and providing feedback to parents (p.4)". CA can be viewed as a contrast to summative external assessment which happens outside the classroom. There are two differences: 1) CA occurs almost on a day-to-day basis whereas external assessment mostly at the end of instruction; 2) CA is a dynamic process with a selection of assessment methods while summative assessments are usually in the form of standard tests (Wang, 2017). CA usually meets the following criteria: 1) The focus of assessment is on documenting individual growth rather than comparing students; 2) the emphasis is on students' strengths rather than weaknesses for improvement; 3) it's an on-going process and is integrated with classroom teaching activities (Angelo & Cross, 2012). The forms of CA include portfolios, self-assessments, peer-assessments, observations, presentations, exhibitions, and journals, written or oral responses and tests can also be applied as a method of classroom assessment (Jiang & Hill, 2018). There is overlap between classroom assessment and formative assessment. Depending on its purposes and functions, the same assessment method can be regarded as classroom assessment and formative assessment. As classroom assessment is conducted during the teaching process, it is mostly formative. Thus, in this thesis, formative assessment is used as an umbrella term that covers classroom assessment.

2.2 Assessment Environment in China and New Zealand

This section describes the assessment environments and national assessment policy in China and New Zealand as contextual background information. As uncontrollable as it is, the external social-cultural educational environment serves as a macro assessment environment that functions as an indispensable contextual factor for assessment conceptions and practices in education (Carless, 2012; Entwistle, 2000; Liu & Xu, 2017).

2.2.1 Testing and its washback effects.

Education and assessment in China and New Zealand vary across a wide range of dimensions. China, deemed as the iconic eastern nation with a Confucian heritage, is known for an exam-oriented assessment environment within a highly centralized education system (Cheng, 2008; Hu, 2005). China is historically known as the country using nationwide examination for selection and promotion (Cheng & Curtis, 2010). Objective testing in China dates back to the Han Dynasty (202 BC – 220 AD) (Spolsky, 1995). The first Civil Service Exam system in the world, *Keju* was originally developed to select people to fill official government positions in the Sui Dynasty (605 AD – 618 AD) (Han & Yang, 2010). Throughout history, exams were used and considered as an essentialtool for the public to upgrade, to achieve and to uplift social status (Cheng & Qi, 2006; Davey, Lian & Higgins, 2007).

In modern-day China, students at school have to participate in numerous exams to succeed in their schooling. The national matriculation examinations of *zhongkao* (provincial entrance exam to secondary schools) and *gaokao* (national entrance examination to college), designed and administrated by provincial and national education bureaus, have the power to decide the fate of almost 70,000,000 test-takers every year (Brown & Gao, 2015).

The national entrance exam in English is a strong stimulus for Chinese English language learners to learn English (Gu, 2014; Liu, 2013; Zhou & Zhou, 2019). However, the negative washback effects of testing in China are clear. Because of the high-stakes of national exams, teachers usually teach to the test and assess with exam-preparation as the guiding principle (Chen & Brown, 2016; Jiang, 2015). Thus, teachers' instruction can be confined to the content being tested rather than a holistic knowledge of the subject being taught (Berry, 2011; Mitchell, 1992; Amrein & Berliner, 2002). Another related problem is that when exams measure the retention of knowledge, rather than application of deep thinking and problem-solving skills, teaching and learning focus on the surface information with repeated, and excessive, practice and drills (Cheng, 2005; Taylor, 2005). Furthermore, the high stakes nature of exams can, and often does, cause great anxiety and mental problems for both high-achievers and low-achievers, affecting their long-term academic and dispositional development (Berry, 2011; Harlen & Deakin, 2003).

2.2.2 China's national policy on English language education (ELE).

Since 2001, China's Ministry of Education (CMoE) has initiated multiple reforms and drafted a series of national education policies to diminish the effects of testing in almost all educational sectors for English language education. The reforms emphasized the importance of using multiple ways of assessment with strong advocacy for formative assessment in ELE (CMoE, 2001; Hu, 2005).

2.2.2.1 ELE policy for primary sector.

For primary English language teaching, the *English Curriculum Standards for Compulsory Education* (CMoE, 2001; 2011) (here after *The Curriculum Standards*) explicitly states that teachers should use formative assessment throughout the teaching process and that the purpose of assessment should be to encourage primary pupils' interest in English study. *The Curriculum Standards* (2011) stated,

> The most important assessment purpose in primary English language teaching is to motivate students' interest and autonomy in English study. The assessment methods should be of a great variety and feasibility with formative assessment as the main. The assessment should be depended on students' interest, attitudes and communicative ability reflected in the daily English teaching activities. The summative assessment can be done in the form of achieved or not achieved or

> Teachers and schools shouldn't use students' academic achievement to rank students as criteria for selection (p. 35).

In addition, *The Curriculum Standards* (2011) specified that there should be no paper-based written tests either in the middle, or at the end, of year 3 and year 4. The assessment for this age group should be conducted during classroom teaching activities through observation and conversation with students. Paper-based written exams can be used for year 5 and year 6 students but should be accompanied by oral exams. The oral exams should measure students' practical ability of using English to communicate.

2.2.2.2 ELE policy for secondary and tertiary sectors.

Likewise, in the secondary and tertiary English language courses, there is a similar promotion of FA. For junior and secondary schools, Chinese Ministry of Education (2001) stated "There should be a focus on formative assessment to help students develop." (p. 7)

> Teachers and schools should establish an assessment system comprising of both formative assessment and summative assessment that motives students' interest in learning and to promote their learning autonomy. The teaching process should be integrated with formative with a focus to cultivate students' confidence and motivation in learning. Summative assessment should focus on evaluating students' comprehensive and practical ability of using English. The assessment should promote students' overall language ability and dispositional development (p. 28).

In the tertiary sector, *The College English Curriculum Requirements* (here after *The Requirements*) was first issued in 2004 for trial, and subsequently revised, with official implementation started in 2007. *The Requirements* stated the principle of assessment as the reform focus, "One of the key aspects for college English curriculum reform lies in assessment. A comprehensive, objective; accurate and scientific assessment system is vital to the curriculum reform" (Chinese Ministry of Education, 2007, p. 18). It further specified the two most fundamental forms of assessment as: "Formative assessment is defined as a progressive and developmental evaluation of students' performance while summative assessment is a conclusive evaluation usually occurred at the end of certain period of study" (Chinese Ministry of Education, 2007, p. 19).

The Requirements strongly advocated establishing an assessment system with variety of assessment tools to develop students' overall English language learning. These policies are interpreted as the guiding principles of the educational reform against the overwhelming influence of testing and examinations and suggest Chinese ESOL teachers employ a greater variety of assessment methods to enhance learning and teaching practice (Li, 2004; Luo, 2003; Wang, 2017).

2.2.3 Assessment policy and environment in New Zealand.

New Zealand schools operate in a pro-formative assessment system within the world's most devolved educational system (Crooks, 2002; 2010). Different from China where each subject has its specific curriculum and assessment requirement, NZ's national curriculum and assessment policy applies to all subjects at all non-tertiary levels. Overall, there's a deliberate focus on the use of professional teacher judgement underpinned by assessment for learning principles rather than a narrow testing regime. The New Zealand Curriculum (New Zealand Ministry of Education, 2007) explains the definition and purpose of assessment as:

> Assessment for the purpose of improving student learning is best understood as an ongoing process that arises out of the interaction between teaching and learning. It involves the focused and timely gathering, analysis, interpretation, and use of information that can provide evidence of student progress. The primary purpose of assessment is to improve students' learning and teachers' teaching as both student and teacher respond to the information that it provides. With this information in mind, schools need to consider how they will gather, analyse, and use assessment information so that it is effective in meeting this purpose.

New Zealand schools are self-managing and function with sufficient freedom to construct and deliver teaching programmes tailored to students' needs (Edwards, 2017). There are no compulsory, large-scale, high-stakes examinations in primary and intermediate sectors (Philips, 2000). In secondary schools there is a standards-referenced national assessment system for entrance to tertiary education known as the National Certificate of Educational Achievement (NCEA). Each of the three levels of the NCEA qualification is achieved through students gaining credits from classroom teacher-marked internal assessment and external assessments (usually summative assessment in the form of end-of-the-year exams or portfolios) (Edwards, 2017). Individual schools and teachers differ as to the range and proportion of the two forms of assessment for NCEA achievement, but nationally there is a higher proportion of NCEA credits gained through internal assessment than external assessment (Crawford, 2016). High stakes testing, while present, is less dominant in the New Zealand educational system than in China. Meanwhile, "Tests and exams in NZ are evaluative for students especially in the final years of schooling, where standardized tests function, for schools as improvement-oriented assessments" (Brown, 2018, p.72).

To conclude, despite the policy advocacy, the purpose of assessment in China is still largely for selection and accountability, thus making assessment in English education highly centralized and examination-focused (Cheng, 2008; Jian & Luo, 2014; Qu & Zhang, 2013). Comparatively, assessment in New Zealand education is espoused as being for improvement and adjustment with lower stakes and less centralized control (Crooks, 2002; East, 2016). The purposes of ESOL assessment, likewise, are mainly to assess how closely students are closing their gap with mainstream cohort and to identify the on-going English language development needs (Franken & McComish, 2003).

2.3 Assessment Literacy in Teacher Education

The concept of literacy has been expanded in recent years. It has been combined with other words to refer to the fundamental knowledge and skills of various domains such as financial literacy, information literacy, and digital literacy among others (Tsagari & Vogt, 2017). Assessment literacy (AL), a synonym for assessment competency in the educational setting, is considered as such an important concept for teachers and educators that it is regarded metaphorically as a “professional suicide” for teachers who neglect it (Popham, 2004, p. 82).

AL is at the core of understanding about assessment preparation in both initial teacher education and teacher professional development across educational systems worldwide (Deluca et al., 2015; Popham, 2013; Volante & Fazio, 2007; Xu & Brown, 2017). Sufficient AL enables teachers to evaluate students’ learning with accuracy and to make valid and reliable decisions, while insufficient AL leads to poor assessment practices which affect teaching and learning (Kane, 2006, Xu & Brown, 2017). From reviewing the literature on AL and teacher education, three aspects were identified for discussion: 1) AL definitions and frameworks for teacher education; 2) AL preparation in teacher education; and 3) the major measurement instruments used to gauge teacher assessment literacy (TAL).

2.3.1 Exploring assessment literacy frameworks for teacher education.

A number of studies have discussed AL topics and language assessment literacy (LAL) by defining the terms and working out a conceptual or practical model for teacher education and training programmes. This section summarizes the major definitions and frameworks used in AL.

2.3.1.1 The knowledge, principles and skills framework.

Stiggins (1995) first raised the concept of AL and stated that an assessment literate teacher would know how, and why, to implement quality assessment activities to maximize learning outcomes. He listed the characteristics of an assessment literate teacher as follows:

- Setting out with a clear purpose;
- Concentrating on the achievement target;
- Choosing an appropriate assessment method (taking the purpose and the target into consideration);
- Judging sampled students achievement accordingly; and
- Minimizing bias and prejudice.

These characteristics focus largely on the skills needed to become assessment literate teachers. Boyles (2006) built his definition of AL by expanding Stiggins' AL definition (1995) emphasizing the practical application and the improvement function of assessment. He suggested that AL is not only the ability to work out and select appropriate assessment methods but entails the capacity to interpret and apply the analysis to promote students' learning, and to "analyze empirical data to improve their instruction without negative repercussions" (p.8). Similarly, other researchers have emphasized connecting assessment theory and practice in an AL definition. Deluca and Klinger (2010) claimed that AL involves "the understanding and appropriate use of assessment practices along with the knowledge of the theoretical and philosophical underpinnings in the measurement of students' learning" (pp. 419-420). Popham (2004) supported this claim, but called for a broader definition and purposes of assessment literacy. They both suggested that a definition of AL should incorporate: the knowledge of theoretical assessment understanding; the principles and the practical ability to apply these principles in teaching which include conducting appropriate assessment practices; interpreting assessment performance; communicating assessment results; and using assessment results to facilitate effective instruction and learning for improvement, accountability or other purposes.

The definition of language assessment literacy (LAL) is similar to the AL definition. Fulcher defined LAL as "The knowledge, skills and abilities required to design, develop, maintain or evaluate largescale standardized and classroom-based tests, familiarity with test processes, and awareness of principles and concepts that guide and underpin practice, including ethics and codes of practice (Fulcher, 2012, p. 125)". Similarly, Berry and O'Sullivan (2016) defined LAL as:

> Skills in identifying and evaluating appropriate assessments for specific purposes within specific contexts; in analysing empirical data in order to improve one's own instructional and assessment practices. The knowledge to assess learners effectively and maximise learning; to interpret and apply assessment results in appropriate ways. And the understanding of the principles and practice of sound assessment. The

> wisdom to be able to integrate assessment and its outcomes into the overall pedagogic process (p.4).

There are numerous frameworks for AL and LAL development in teacher education based on these definitions. Deluca established a conceptual framework for assessment education in a three-tiered learning model (i.e. I.C.E Model). The first tier I represents “Idea”, which requires a student teacher (ST) to get involved in explicit teaching and learning of fundamental concepts and principles to assessment. The second tier, C means “Connections”, which encourages a ST to connect the fundamental ideas together and relate these ideas to their personal reflections and experiences. The final tier, E means “Extension” and refers to the application or extension of the ideas and the connections to aspects of their teaching practices (Deluca & Lam, 2014). The Siegel and Wissehr’s (2011) framework looks similar to the I.C.E model but emphasizes the alignment between theoretical assessment principles, specific assessment purposes and assessment tools with classroom practices.

Davis (2008), likewise, proposed a framework of knowledge + principles + skills for teaching LAL to language teachers, emphasizing that the ultimate goal for assessment education should be learning skills through specifically designed activities, and should be supported by the knowledge one acquires from this process to construct a context for skills. The principles direct the appropriateness, quality and the effectiveness of the language assessment, as Davies (2008) explained in his work:

> Skills provide the training in necessary and appropriate methodology, including item writing, statistics, test analysis and increasingly software programmes for test delivery, analysis and reportage. Knowledge offers relevant background in measurement and language description as well as in context setting. Principles concern the proper use of language tests, their fairness and impact, including questions of ethics and professionalism (p. 58).

2.3.1.2 Increasing attention to formative classroom assessment.

Aligned with the shift from summative standardized testing to formative assessment, as discussed in section 2.1.2.3, formative and classroom assessment has received considerable attention in teacher education. Brookhart (2003) and Hill et al. (2010) called for an increase of teaching of formative classroom assessment in teacher education programmes. Kane (2006) argued that teachers should not only be able to interpret test scores but also make sound and reasonable interpretations of students’ performance using various forms of assessment methods. Inbar-Lourie (2008) agreed by stating that language assessment training should “give due emphases to both classroom and external assessment practices” (pp. 396-397).

Berger (2012), building on the evolving concept and discussions of AL, suggested that the content of language assessment programmes should focus not only on large-scale testing but also on classroom-based assessment. He argued that summative testing and formative assessment methodologies should be treated equally. He proposed a three-dimensional conceptual model for teacher education that can be used to “operationalize the concept of language assessment literacy” (p. 72) suitable for different populations and purposes. The model emphasizes that instruction should embed reality into classroom teaching, contextualize teaching in the subject taught and provide examples and specific guidance on formative assessment methods, rather than just the popular ones. Central to this idea is that the programme designer should have a broader view of language assessment than just the testing approach, and the instruction should be expanded to include the concept and teaching of classroom-based assessment (Berger, 2012).

2.3.1.3 Considering the social-educational context.

Acknowledging the social dimension of assessment, several researchers argued that both AL and LAL should be considered within the wider social-cultural contexts (e.g., Inbar-Lourie, 2008; Fulcher, 2012; Malone, 2013; Taylor, 2009; Xu & Brown, 2017). Inbar-Lourie (2008) suggested positioning teacher LAL development in a social-constructivist perspective and promoting a holistic and dynamic understanding of assessment outside the language testing culture and community. The major components for assessment literacy development that language teachers acquire in assessment training should, Inbar-Lourie argued, “reflect current views about the social role of assessment in general and language assessment in particular, contemporary views about the nature of language knowledge" (pp. 396-397).

Fulcher (2012) supported the social dimension of AL development when reporting on a survey regarding a needs-analysis of language assessment training. He proposed that an assessment literate teacher should place knowledge and skills in the wider external environment.

> The ability to place knowledge, skills, processes, principles and concepts within wider historical, social, political and philosophical frameworks in order to understand why practices have arisen as they have, and to evaluate the role and impact of testing on society, institutions, and individuals (p. 125).

Xu and Brown (2016) constructed a conceptual framework of teacher assessment literacy in practice (TALiP) through a scoping review of AL studies in educational assessment teacher education. Their pyramid-shaped framework consisted of six components from the bottom to the top level: the knowledge base; teachers' conceptions of assessment; institutional and social-cultural contexts; TALiP as compromises made among tensions; teacher learning; and teacher identity re(construction) as assessors. This comprehensive framework integrates previous frameworks and includes most of the key components of teacher assessment literacy (TAL).

To summarize, the definition and framework of AL, LAL has been developed and broadened by researchers in the last decades by considering the influential and mediating factors in the wider social-cultural and educational contexts. The knowledge + principle + skill definition and framework appears as the most fundamental and essential framework. There's increasing attention to formative classroom-based assessment in AL framework and definition, thus it is recommended that AL preparation in teacher education should be set in wider social-cultural and educational context.

2.3.2 Enhancing assessment literacy in teacher education programmes.

There are a number of studies on AL preparation in general teacher education (e.g., Brown et al., 2011; Deluca et al., 2013; Eyers, 2014; Hill et al., 2010) as well as in language teacher education (e.g., Malone, 2013; Fulcher, 2012;). This section summarizes the assessment preparation in general teacher education programmes and LTACs in TESOL programmes respectively.

2.3.2.1 Assessment preparation in general teacher education programmes.

Aligned with the increasing focus of formative assessment classroom, researchers have suggested specific contents should be included to make assessment more related to classroom teaching (Airasian, 1991; Stiggins, 1991; Mertler, 2004). These researchers advocated expanding courses from testing principles and techniques to formative classroom assessment practices aligned to curriculum requirements of teaching and assessment to help teachers acquire a higher level of AL "to meet the challenge of day-to-day classroom assessment" (Stiggins, 2002, p. 760). However, the studies also indicated the challenges of assessment learning during the preparation process, which reflected a tension between summative assessment and formative classroom assessment.

From 2010 to 2012 in New Zealand, a longitudinal study on the assessment preparation for pre-service teachers at four New Zealand universities were carried out with the purpose to understand how teachers (Hill et al., 2010).Teacher educators in assessment courses were interviewed, and surveys and focus group discussions with preservice teachers were conducted. Analysis of the data showed changes among pre-service teachers in their assessment beliefs during three years of study. Pre-service teachers, prior to commencing teacher education, regarded assessment as having primarily a summative purpose with testing as the major form. Upon graduation, their opinions about assessment were broadened which included alternatives in assessment and acknowledged the formative purpose and process. The assessment preparation courses and practicum are reported to be the two most important factors that help enhance pre-service teachers' assessment ability.

Mertler (2009) designed a curriculum with "The Standards Project" as the guiding principles in an initial teacher education programme in America and found it successfully enhanced pre-service teachers' understanding of assessment as a systematic and formative process. However, Siegel and Wissehr (2011) identified a gap between pre-service teachers' conceptions about assessment, and their practical assessment planning in their investigation of the development of the assessment literacy for 11 secondary science teachers during their pre-service teacher education. While the conceptions they describe about assessment were formative, their planning for assessment was more testing centered. They acknowledged the need to align assessments with learning objectives, as well as the importance of using multiple forms of assessment to improve learning. Their unit teaching plans, however, did not include formative assessment methods, instead they resorted to traditional ways of testing and grading to assess students' learning.

Similarly, Moss (2013) noted that when graduating, pre-service teachers were confident, but less competent in using assessment to improve student learning. She suggested that researchers needed to focus more on "examining what actually happens in those courses to develop assessment literacy and follow graduates into the field to see if those courses impact actual assessment practices" (Moss, 2013, p.252). Deneen and Brown (2016) investigated the impact of teaching the concepts of assessment on AL development in a post-graduate level educational assessment course with both pre-service and practicing teachers. The course was constructively designed to link teaching, learning and assessment (Biggs, 1996). Its aims were to enhance AL and the outcomes were based on students' achievement relevant to these objectives. Using a mixed research method, they found that student teachers were very concerned with the consequences of assessment and its fairness.

To conclude, these studies found that teacher education programmes with well-structured and well-designed assessment courses or preparation processes can help improve students' assessment literacy. However, the constraints within initial teacher education programmes suggest that only a rudimentary level of assessment literacy can be expected of pre-service teachers. In light of this, researchers internationally argue for a closer scrutiny of what, and how, assessment preparation courses can best enhance in-service teacher education and support ongoing teacher assessment literacy development (Edwards, 2017, Black et al., 2010). As Hill et al. (2010) claimed in their review of the literature on assessment education:

> Overall, there appears to be meagre systematic evidence internationally about how student teachers are successfully prepared in assessment, their readiness to use effective assessment strategies after graduate, and how they learn to use the assessment practices know to improve student outcomes. (p. 55)

2.3.2.2 LTACs in TESOL programmes.

Studies on LTACs have stepped into spotlight in recent years for educational, career and social political reasons (Taylor, 2009). LTACs are taught mostly in graduate and postgraduate programmes particularly in the field of TESOL or Applied Linguistics, either as a selective or compulsory course. However, compared with the ample research on other aspects of language assessment and language teacher education, studies about LTACs in the preparation of ESOL teachers is still in its infancy (Brindley, 2001). An analysis of studies on LTACs in teacher education programmes raises the following concerns: 1) The course focus was predominantly on language testing; 2) the course instructors have tried to reform their courses by putting greater focus on practice.

2.3.2.2.1 Course focus on language testing.

Nearly three decades ago Bailey and Brown (1998) sent out questionnaires (with both Likert-scale and open-ended items) to instructors through the mailing list of Language Testing Research Colloquium to examine the background of instructors, course content and student teachers' attitudes towards the course. Their study reported that statistics in testing, item-writing, and test development were the main aspects covered in LTACs. A decade later, adhering to the same research question and methods but with 20 more items added to the original questionnaire, Brown and Bailey (2008) again set out to investigate the emphases of language testing courses and discovered that little had changed.

An acknowledged weakness of Bailey and Brown's (1998) research is that it was instructors from the United States who mainly responded to the questionnaires. Other researchers have argued for the importance of investigating courses in different cultural contexts. Using similar methodology, Jin (2010) conducted a national survey with over 80 language assessment and testing course instructors in China. The focus was on their teaching content, teaching methodology, student perceptions of the courses and teaching materials. Instead of focusing exclusively on language testing courses, Jin included courses entitled 'language assessment'. Her survey revealed that the theory and practice of language testing was a priority with little attention given to educational and psychological measurement, and even less to classroom assessment practices. In a similar context, Lam (2015) described the overall landscape of assessment training of pre-service ESOL teachers by investigating five teacher education institutions in Hong Kong against the backdrop of the assessment reform. He discovered that despite strong advocacy of assessment for learning by the Hong Kong government, language assessment training in Hong Kong was inadequate; many of the courses were about theory and technical skills of test construction and test measurement. There was little support for student teachers' language assessment literacy development.

Although the concept of assessment covers both out-of-classroom summative testing and in-classroom assessment activities, most studies to date have shown that an overwhelming amount of teaching and learning LTACs focuses on the knowledge about, and training for, language testing for pre-service ESOL teachers with comparatively little attention to assessment as a more comprehensive concept. With a number of the courses entitled language testing, the teaching content is weighted largely towards test construction, test analysis and other test related issues (Brindley, 2001; Jin, 2010). This means, inevitably, the content of assessment is diminished.

As Malone (2008) pointed out, a major consideration in language assessment training is the construction of items for standardized tests rather than teaching prospective teachers how to assess their students. A possible explanation for the preference for testing in language teaching may be that it takes so much time to be a professional in language testing teachers while the students have little time for other potentially rewarding subjects (Davies, 2008). Furthermore, traditional ESOL teaching and learning used testing as the major form of assessment (Bachman, 2004; Brindley, 2001) and ESOL teachers are involved in testing early in their career through devising tests for students, and classifying students according to test results in the entrance exams (Wharton, 1998). A more specific reason is the overwhelming influence of the testing culture that dominates the educational system in countries (Berry, 2011; Gu, 2014; Shohamy, 2007).

There is growing evidence, however, that assessment training has changed over time to reflect trends towards the use of multiple assessment methods in language assessment. As argued by Stiggins (2002), the overall effectiveness of assessment is diminished if it does not work well in everyday classroom, which highlights the importance of classroom-based assessment. The growing professionalism in language testing has two problems: 1) The large amount of time required to learn test construction and theory constrains time for students to become proficient in classroom-based assessment (Davies, 2008); 2) the comprehensive learning resources over protect students from getting contact with empirical practice with real language learners. Due to that, most of the learning is book-focused and there is little emphasis on the practice with real actual learners (Graham, 2005).

2.3.2.2.2 Course instructors' attempts to reform the course.

Studies on LTACs focused on the background information about the instructors (Brown & Bailey, 2008; Jin, 2010; Lam, 2015), but their voices and attitudes towards assessment were less documented. However, great challenges confronted course instructors when they have to integrate the complex disciplinary content into the course and facilitate students' learning (Wharton, 1998). Most course instructors have to deal with limited time in preparing ESOL teachers, which means that the course content has to be compressed (Kleinsasser, 2005). The main teaching method reported by course instructors, therefore, was lecture-based and transmitting knowledge to maximize content delivery (Singh & Richards, 2006). Inbar-Lourie (2008) advocated that language assessment professionals should contribute to this field through a reform in pedagogical and assessment methods. Graham (2005), Kleinsasser (2005) and Wharton (1998) reported three cases in which the instructors applied transformative teaching of language assessment instead of the traditional knowledge transmission approach.

Kleinsasser (2005) described a narrative inquiry technique to document and analyze his transformation of pedagogical method from a content-based to a learner-centered instruction approach in a postgraduate TESOL (Teaching English to Speakers of Other Languages) testing and assessment course. He combined students' learning with professional development in assessment by negotiating the assessment contents and assessment material development with students, and using portfolios for assessment tasks. Students' learning was intended, therefore, as professional development rather than a by-product of professional development. Kleinsasser (2005) found students showed greater autonomy in learning, with the final course evaluation indicating students' preference for this form of teaching. Wharton (1998), similarly, assigned multiple tasks in her language assessment courses: she involved students in the test design and evaluation rather than relying on the lecture-based teaching method. She claimed success for this task-based learning pedagogy for the following reason: Student teachers were exposed to various assessment formats and procedures; the process of test design and construction was better conceptualized; and it was a more practical application of language testing theory. This student-based and task-based pedagogical method was replicated and supported by Johnson, Becker and Olive (1999). It was further developed to focus not only on the test construction from the test givers' perspective, but also on test analysis and how to generate useful information from the test results.

In conclusion, course instructors acknowledged the importance of students' involvement in the assessment preparation courses and that their shift of instruction from a solely knowledge-transmitting way to a more practical interactive approach brought about fruitful changes. The teaching content, or the focus of the LTACs, however, still appeared to be weighted towards language testing. The challenges for the courses come from the limited time and attention devoted to LTACs in teacher education programmes (Taylor, 2009). Challenges in preparing for courses and a lack of assistance in resolving difficulties are constantly reported (Kleinsasser, 2005). In Wharton's (1998) course, they had only four weeks to deliver all the main classes while in Johnson's (2010) course they had two months. Furthermore, although there is a range of textbooks on language testing and assessment available today, many are highly technical, or too specialized, for the course instructors to use (Davies, 2008).

Finally, language assessment preparation courses are weighted largely towards language assessment as testing with less weighting given to classroom-based assessment. Some instructors recognize the importance of involving students in the process of learning to design and comment on assessment tasks, but mainstream pedagogical methods adopted appear to be still lecture-based knowledge transmitting.

2.3.3 Measuring teacher assessment literacy.

Another important aspect of studies on AL preparation in teacher education programmes involves the measurement of AL. Several researchers have developed instruments in an attempt to measure TAL (e.g., DeLuca, LaPointe-McEwan, & Luhanga, 2016; Plake, 1993; Mertler, 2004; 2005). Some instruments have been based on national and international teacher qualification standards, while others have measured teacher assessment literacy by focusing on conceptions and practices of assessment.

2.3.3.1 Measurements based on The Standards.

In the beginning of the 1990s, The American Federation of Teachers (AFT), the National Education Association (NEA) and the National Council on Measurement in Education (NCME) (1990) jointly drafted *The Standards for Teacher Competence in the Educational Assessment of Students* (*The Standards*) to set a yardstick for teacher assessment competence so that the potential benefits of assessments can be fully realized. *The Standards* described assessment as “a process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths, and weakness, to judge instructional effectiveness and curricular adequacy, and to inform policy” (AFT, NEA & NCME, 1990, p.30-32). Based on this definition, a set of seven key competencies for an assessment literate teacher was established which specified that teachers should be skilled in:

1. choosing assessment methods appropriate for instructional decisions;
2. developing assessment methods;
3. administering, scoring and interpreting the results of both externally produced and teacher-produced assessment methods;
4. using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement;
5. developing valid pupil grading procedures that use pupil assessments;
6. communicating assessment results to students, parents, other lay audiences and other educators;
7. recognizing unethical, illegal and otherwise inappropriate assessment methods and uses of assessment information (AFT, NEA & NCME, 1990, pp.30-32).

The Standards became the basis for a series of teacher assessment literacy instruments worldwide. Among these instruments, the first and most widely used was the Teacher Assessment Literacy Questionnaire (TALQ) (Plake, Impara & Fager, 1993). TALQ consisted of 35 multiple-choice items composed of five classroom-based scenarios, each containing seven items mapped to the seven competencies proposed in *The Standards*. The items were developed as application-type questions that reflected teachers' practical teaching. TALQ was later reworded and revised (e.g. the teachers' names have been changed, some sentences have been rewritten for clarity and seven demographic items were added) by Mertler (2004; 2005) with a new name called Classroom Assessment Literacy Inventory (CALI), but later renamed as ALI into a revised version (Mertler, 2009; Mertler & Campbell, 2005). Table 1 lists a summary of studies using *The Standards*-based AL instruments.

Table 1

Studies on AL Measurements

Authors/Year	Instrument	Participants	Findings
Plake (1993); Plake, Impara, & Fager, 1993)	TALQ	555 non-tertiary in-service U.S. teachers	The reliability was below the acceptable threshold; Teachers lack AL as reflected from the low scores.
Mertler (2004)	CALI	67 pre-service teachers and 101 in-service teachers	In-service teachers scored higher than pre-service teachers.
Mertler & Campbell (2005)	ALI (revised)	249 pre-service teachers	The scores were not satisfactory indicating lack of AL.
Mertler (2009)	ALI (revised)	7 in-service teachers	In-service teachers' AL increased dramatically in the specifically designed two-week assessment course.
Chen (2005)	TALQ	25 undergraduate and 36 graduate teacher candidates	Graduate teachers had higher AL; Previous teaching experiences had positive association with AL.
Alkharusi (2011a)	TALQ	516 Oman in-service teachers	Attitude, self-efficacy, assessment training and years of teaching experiences are reliable predictors of AL
Alkharusi (2011b)	TALQ	259 Oman pre-service teachers	A technical analysis. Both the items and the constructs of TALQ showed acceptable validity and reliability.
Alkharusi, Kazem and Al-Musawai (2011)	TALQ	279 pre-service teachers and 233 in-service teachers in Oman	In-service teachers had higher level of AL, more favourable attitudes to assessment and a higher level of confidence.
McGee & Colby (2014)	ALI	190 pre-service teachers	Assessment course improves AL.
Hailaya, Alagumalai, &	ALI	582 pre-service Philippines teachers	A technical analysis. ALI works well at item level but rejected the seven-

Ben, 2014			factor model based on the seven competencies.
Beziat & Coleman (2015)	TALQ	Pre-service teachers, 49 at the beginning, 26 at the end	Pre-service teachers lack AL despite completing assessment course; Student teachers with previous learning on assessment showed greater AL.
Xu & Brown (2017)	TALQ (adapted)	891 Chinese tertiary EFL teachers	A technical analysis. Validity for limited number of items; The multi-dimensional model was rejected.

Note. Studies presented in chronical order.

The above-mentioned measurement instruments based on *The Standards* have several limitations: 1) the questions are mostly context driven, being designed for the North America; 2) the scenarios contain some terminology and may be too difficult for pre-service and novice teachers to understand and 3) it only focuses on the content knowledge and perceived skills of assessment.

A synthesis and analysis of these studies identified that: 1) Despite its importance, teachers' assessment literacy was inadequate, 2) assessment course in general helps enhance AL level except in Beziat and Coleman (2015)'s research that pre-service teachers showed a lack of AL despite completing assessment course and 3) in-service teachers performed at a higher level of AL than pre-service teachers, suggesting that classroom teaching experience enhances AL level; 4) confidence or also called self-efficacy for assessment is a contributing variable to high levels of AL.

Self-efficacy refers to one's belief in his or her ability to execute behaviors needed for specific tasks or attainments (Bandura, 1997; 2010). Self-efficacy level can determine one's resolution and ability to cope with obstacles and implement the task (Maibach & Murphy, 1995; Stajkovic & Luthans, 1998). Self-efficacy of assessment is an important factor for teachers to apply effective and appropriate assessment and thus should be considered as an important component to assessment (Guo et al., 2012). It is estimated that high level self-efficacy for assessment help teachers to make consistent and sufficient efforts that lead to successful outcomes whereas low-self efficacy hinders individual's assessment practice.

2.3.3.2 Measurements on conceptions and practices of assessment.

The effectiveness of assessment preparation in teacher education programmes is determined by how teachers conceive of assessment and how effectively they implement assessment literacy in practice. Thus, the conceptions and practices of assessment are important determinants of teachers' assessment literacy development (Deneen & Brown, 2016). Within the context of this research project, the term "conceptions of assessment' is inclusive of the beliefs, attitudes, and perceptions one holds about assessment, that is, "the systematic frameworks for understanding assessment and that includes people's attitudes towards it" (Brown, Irving & Keegan, 2014, p.5); "practices of assessment" refers to the assessment plans, tasks, and activities conducted by teachers.

2.3.3.2.1 Measuring conceptions with TCoA.

The most prevalently used instrument to measure teachers' conceptions of assessment is the Teachers' Conceptions of Assessment (TCoA) inventory created by Brown (Brown, 2002; Brown, 2004). TCoA initially contained 50 items but later was abridged to 27 items using a positively packed 6-point scale (Brown, 2008). TCoA measures how teachers conceive the nature and purposes of assessment in four inter-related conceptions: 1) Assessment makes students accountable for their learning (student accountability); 2) assessment makes schools and teachers accountable for their effectiveness (school accountability); 3) assessment improves teaching and learning (improvement); and 4) assessment is irrelevant (irrelevance). The first three are termed as "purpose" while the last one as an "anti-purpose" (Brown, 2008; 2011).

Brown and his colleagues have conducted several studies on primary and secondary school teachers' conceptions of assessment in the New Zealand context (Brown, 2002; 2004). The NZ primary school teachers on average agreed that: 1) Assessment was for improvement of learning; and 2) teachers and schools should be accountable for assessment, but rejected that assessment was irrelevant and used for student accountability. Brown (2008) also investigated the link between assessment literacy training and NZ primary school teachers' conceptions of assessment. The results showed that, increased levels of assessment training were not associated with higher acceptance for the improvement purposes of assessment, nor were they associated with less acceptance of irrelevance and accountability purposes.

Brown (2011) further surveyed pre-service teachers' conceptions of assessment and compared these with their in-service counterparts. The findings indicated pre-service teachers differed substantially from practicing teachers in terms of their conceptions of assessment. More pre-service teachers agreed that assessment is for the purpose of accountability while the majority of in-service teachers indicated that assessment should be for improvement purposes. TCoA was used also in Queensland to survey Australian primary and secondary school teachers' conceptions of assessment; the findings showed that Queensland teachers held similar conceptions as compared with their NZ counterparts. However, in Queensland, more primary than secondary teachers agreed assessment had an improvement purpose whereas more secondary teachers agreed that assessment was for student accountability (Brown, Lake & Matters, 2011).

Brown (2008) claimed that "what we believe about the nature and purpose of something influences what we do around that something and our practices determine, in part, the outcomes" (p.16). Empirical studies on the relationship between conceptions and practices of assessment have explained how conceptions and practices interact with each other. Brown's (2011) study with New Zealand primary school teachers indicated that when teachers conceive assessment as accountable for students' learning, they are more inclined to choose the formal and test-like assessment practices. In contrast, when assessment is viewed as an indicator of school quality, it is more likely that teachers will choose to use measures of deep learning. The conception that assessment is for improvement was linked positively with formative assessment practices such as student-teacher interactions and self-/peer assessment.

2.3.3.2.2 TCoA in the Chinese context.

Differences in culture or society lead to distinctive conceptions and practices of assessment. Brown, Lake and Matters (2011) argued that "conceptions are ecologically rational representations of the thought and practice traditions an individual experiences within a culture" (p.308). New Zealand and Queensland share similar low-stakes and pro-formative assessment environments and teachers' conceptions and practices are relatively similar. In contrast, in strong examination-oriented educational systems, teachers' conceptions and practices showed different results. A survey using an Arabic version of TCoA with 507 Egyptian schoolteachers indicated a strong positive relationship between Improvement and Student accountability (Gebril & Brown, 2014). In India, it was found that greater use of diagnostic practices depends in part on teachers believing in the positive role of internal, school-based assessment and that educational improvement, as the legitimate purpose of assessment, is to be encouraged (Brown, Chaudhry & Dhamija, 2015).

There are several studies pertaining to the Chinese context where educational assessment serves as an effective tool for improvement of the assessment results. In China’s test-driven assessment environment, Chinese teachers view assessment in context-specific ways, thus the TCoA inventory was revised for the specific Chinese assessment environment. Li & Hui (2007) translated the TCoA into a Chinese version of TCoA and surveyed Chinese colleague teachers. The results suggested Chinese teachers agreed that assessment could improve quality of teaching and learning but also made school accountable. They expressed doubt, however, whether assessment could provide reliable information and used for improvement of learning in contrast to passing exams. Brown et al. (2009) conducted another survey, with Hong Kong primary teachers, using two inventories, the Chinese translation of TCoA and a Practices of Assessment Inventory to measure the relationship between conceptions and practices and the results showed that: 1) Hong Kong teachers strongly associated the improvement purpose of assessment with making students accountable, and that these beliefs lead to teachers’ examination preparation practices; 2) Hong Kong teachers believed assessment should be used to make students accountable and, together with exam preparation practices, students’ learning outcomes could be improved.

These studies indicate that the original TCoA only tapped into part of Chinese teachers’ conceptions of assessment. Due to the specific exam-driven assessment environment, Chinese teachers conceived accountability quite differently from their NZ and Queensland counterparts. A TCoA inventory for the Chinese context (C-TCoA) was developed, therefore, based on previous studies (Brown et al., 2011) and related qualitative studies on Chinese teachers’ conceptions of assessment (Li & Hui, 2007).

It was found that Chinese teachers believed assessment could change students’ attitudes of learning, could identify their potential and prepare students for future challenges, which were consistent with the Chinese teachers' idea that high assessment results indicate a better person (Chen & Brown, 2013). The study using C-TCoA showed that the Chinese teachers strongly associated accountability with improvement (Brown et al., 2011). Brown and Gao's (2015) article further explained the six competing and complementary conceptions of assessment held by practicing Chinese teachers. From the positive perspective they regarded assessment as developing personal qualities and academic abilities of students, from a negative perspective they viewed assessment as management and inspection of schools.

In conclusion, studies using TCoA and its different versions showed that: 1) Teachers' conceptions of assessment were consistent with the policies and priorities of the educational environment in which teachers studied and worked; 2) teachers' conceptions and practices should be measured with instruments specific to the social-cultural and educational context; and 3) teachers' conceptions of assessment, theoretically and empirically, largely influenced their assessment practices (Kahn, 2000).

2.4 Chapter Summary and Research Questions

Assessment in this thesis is defined as both a method and process to gather students' information for the purposes of selection, improvement and qualifications. Language assessment has its unique disciplinary characteristics that it has a communicative function and social dimensions. Assessment in both general education and language education has seen a shift of focus from summative standardized testing towards the inclusion of classroom based formative assessment.

The definition of assessment literacy and language assessment literacy is expanding. Researchers called for the incorporation of classroom assessment into assessment preparation courses and a practical teaching approach. With the review of the major AL measurements, it was found that the AL measurements based on *The Standards* focused on the knowledge and skills of assessment while other measurements focused on teachers' conceptions and practices of assessment. Considering the importance of self-efficacy in assessment as discussed in section 2.3.3.1, it is thus proposed that self-efficacy, conceptions and practices of assessment be considered as three important components to measure student teachers' assessment literacy for this thesis.

The literature review, therefore, informs the overarching research questions of this research as follows: "How are LTACs constructed and implemented in different cultural and academic contexts, and how do they shape students' assessment literacy". Sub-questions to guide the research are:

1. How are LTACs constructed and implemented under different cultural and academic contexts?
2. What are the commonalities and differences of LTAC construction and implementation in these contexts?
3. How do LTACs help develop student teachers' assessment literacy?

Chapter 3 explains and justifies the methodology employed in addressing these questions.

Chapter 3 Methodology

Research methodology is the systematic and theoretically guided approach taken to explore a research question (Creswell, 2013). According to Creswell and Clark (2017), methodology includes the principles of, and procedures for, research design, data collection, data analysis, and data interpretation. Typically, it encompasses explanations about such things as paradigm, theoretical model, and research techniques. Based on the literature review and the research questions raised in Chapter 2, this chapter explains the methodology employed in this thesis including sections on research paradigm and methods, research design, instrumentation, data collection, data analysis and ethical considerations.

3.1 Research Paradigm and Methods

Either consciously or subconsciously, we view the world with an embedded set of beliefs, based on which we analyze our relationship with the world and interpret what we observe. This ontological, epistemological and methodological belief system is known as a paradigm (Guba & Lincoln, 1994). As Guba & Lincoln wrote, "Questions of method are secondary to questions of paradigm" (Guba & Lincoln, 1994, p.105). The idea that a researcher can set aside his or her own background, knowledge, experience, and theoretical learnings when embarking on research, and play the role of passive and objective observer seems daunting and implausible (Johnson & Onwuegbuzie, 2004). Hence, a research paradigm should be placed with the foremost priority for researchers before they embark on their research journey.

This thesis investigated the construction and implementation of language testing and assessment courses (LTACs) in different contexts to probe into how contextual factors impact the courses and how such courses help develop student teachers' assessment literacy. As the research questions stem from practical teaching and learning inquiry, and the research set out to use empirical studies to elicit answers that have practical implications, a pragmatist paradigm (Morgan, 2014) was adopted. Within this paradigm, the desire is to probe into "the singular and multiple realities that are open to empirical inquiry and orients itself toward solving practical problems in the real world" (Feilzer, 2010, p.8). Meanwhile, a pragmatist approach offers strong support for work that uses mixed methods (Morgan, 2014) for it lends itself well to the mixing of qualitative and quantitative methods to gain a practical understanding of a phenomenon (Colapietro, 2006).

Mixed methods research is "an approach to research in the social, behavioral, and health sciences in which the investigator gathers both quantitative (close-ended) and qualitative (open-ended) data, integrates the two, and then draws interpretations based on the combined strengths of both set of data to understand a research problem" (Creswell, 2014, p. 2). Researchers view mixing qualitative methods and quantitative methods of various data sources as not only feasible but desirable (Teddlie & Tashakkori, 2009). Mixed methods research methods were employed in this thesis because it can provide a more comprehensive understanding of research issues than either qualitative or quantitative methods alone (Palinkas et al. 2011). Qualitative methods enable a profound and thorough understanding of detailed issues around the LTACs construction and implementation while the quantitative methods are used to test the assumptions on the conceptual model of students' assessment literacy and the effect of LTACs. The combination of the two research approaches can enhance research reliability, validity and trustworthiness with rich and in-depth textual as well as statistical information (Watkins & Gioia, 2015). Moreover, mixing methods helps to uncover new knowledge about complicated and multifaceted phenomena from different perspectives and provide practical implications from the collected data (Yin, 2010). Thus, in addition to examining data from literature and theory already discussed, this thesis employs a pragmatism research paradigm with mixed methods research.

3.2 Research Design

This thesis is an exploratory mixed-methods design that triangulates three types of data (course syllabi content analysis, course instructors' interviews and students' survey responses) in which each type of data was analyzed with different methods. There were two sequential phases of data collection. Phase-I qualitative data included course syllabi and course instructors' interviews from China (at both undergraduate and postgraduate levels) and New Zealand (only postgraduate level available). Phase-II quantitative data collection gathered Chinese senior undergraduate students' (the ESOL pre-service teachers) responses to a questionnaire. Three studies were designed based on the data collected: Study 1 used content analysis to investigate the course syllabi data; Study 2 examined interview data using thematic analysis. Building on insights gained from Study 1 and Study 2, Study 3 examined students' survey responses and analyzed data mainly using statistical analysis methods of confirmatory factor analysis (CFA) and structural equation modeling (SEM). In addition, this thesis uses a comparative lens to compare data from New Zealand with data from China, and data from Chinese undergraduate courses with Chinese postgraduate data. These comparisons are evident in Studies 1 and 2 and provide a framework through which the Chinese student survey in study 3 can be interpreted. Figure 1 visually presents the research design.

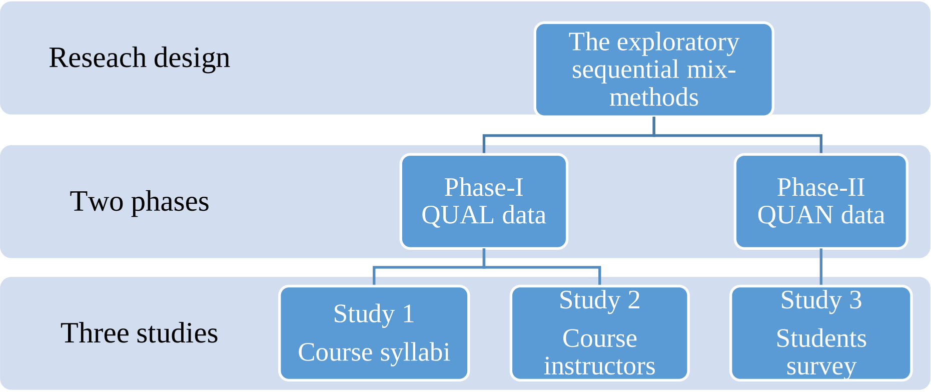

Figure 1. The exploratory mixed-methods design for this thesis.

Note. QUAL = qualitative; QUAN = quantitative.

3.3 Instrumentation

The concept of instrumentation in this section has been expanded to refer to the tools used to examine or measure the research subject in this thesis. This section describes the instruments used in each study: Study 1, an evaluation framework; Study 2, the interview protocol and Study 3, the Conception, Self-efficacy and Practice of Assessment (CSEPoA) questionnaire.

3.3.1 Evaluation framework for course syllabi.

In institutes of higher education, where words and texts serve as a main source of information to communicate academic thoughts and ideas, the course syllabus, along with the lecturers and facilities, is an indispensable component. A syllabus, the pre-set and prescribed documented course information, functions as the official and fundamental source of information for both students and teachers. Graves, Hyland and Samuels (2010) metaphorically consider the syllabus acts as a contract between students and instructors that mediates their relationship. Although a course syllabus may not contain all the course details, this written record generally provides a fairly accurate representation of the aim and scope of the course (Stanny, Gonzalez & McGowan, 2015). Course syllabi serve as an essential indicator of the core purpose, content and assessment scheme (Comeaux, Brown & Sieben, 2015), and form an important part of course design and construction. Also, as the “unobtrusive but powerful indicators of what takes place in classrooms” (Stanly et al, 2015, p.159), a course syllabus functions as an individualized representation of the lecturers’ idealized picture of the course through which their best intentions and practice can be described.

A course syllabus usually contains a great amount of information. Slattery and Calson (2005) proposed that the best practice for preparing an effective syllabus should include “course goals and objectives, means to achieve these objectives, methods of grading and a schedule of events (p.160).” Similarly, Stanny et al. (2015) created a syllabus review rubric that listed the most essential components of a syllabus as: class meeting time and location;

required textbooks and readings; description of course; course goals or learning outcomes; course content; and, grading rubrics for assignments and assessment. O'Brien, Millis, and Cohen (2009) listed learning outcomes, assessment schemes, textbook and content topics as the essential components of a course syllabus evaluation. Based on these studies, four key components of course syllabi were treated as focus of analysis in this study: namely, general course information, course objectives, course content and assessment plans. Each was analyzed with an analysis checklist. In addition, as the research question examines LTACs in different cultural and academic contexts, "national differences" (between China and New Zealand) and academic levels (at undergraduate and postgraduate levels) were used as categories to divide LTACs into three groups for comparison: Chinese undergraduate, Chinese postgraduate and New Zealand postgraduate LTACs. In addition, Davis (2008)'s knowledge + principle + skill assessment literacy framework (see section 2.3.1.1, Chapter 2 for more details) was adopted as the theoretical framework for analysis. Taking these concerns into consideration, an evaluation framework was developed for Study 1 analysis, which is presented in Figure 2.

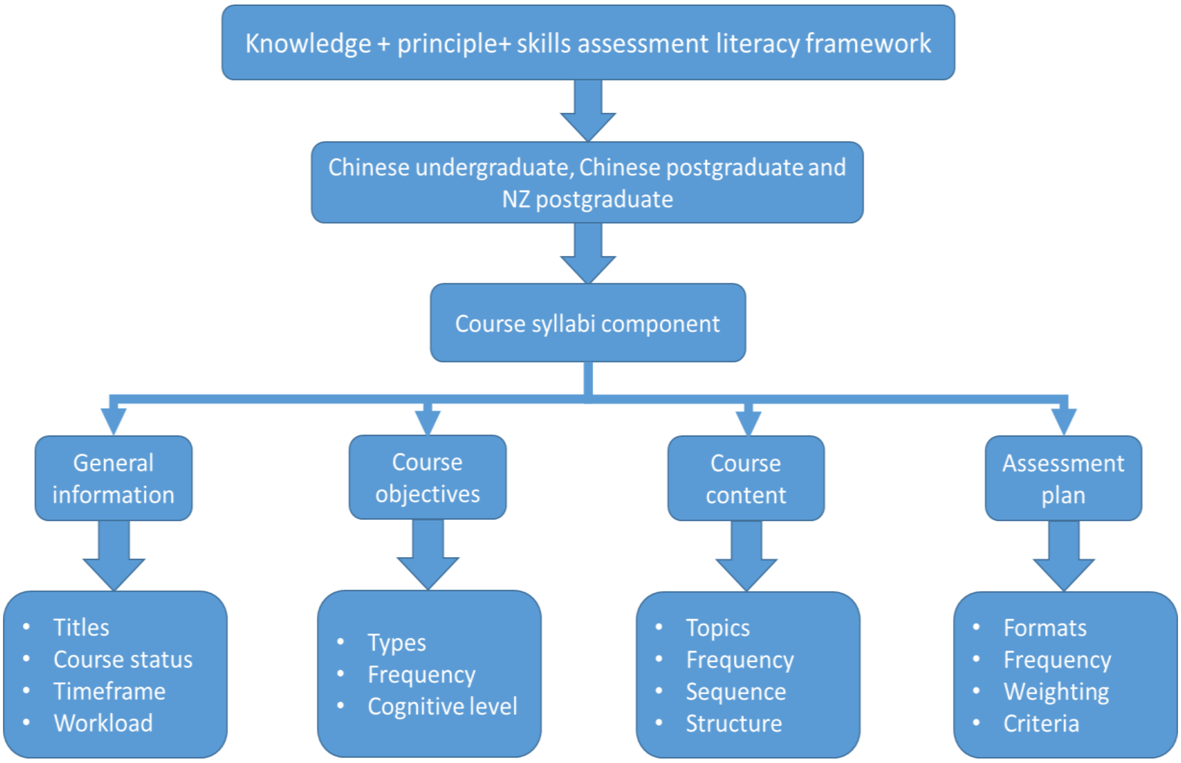

Figure 2. Evaluation framework of course syllabi for Study 1.

3.3.2 Interview protocol for course instructors.

Course instructors, the teachers or lecturers who teach the course and facilitate student learning during the course, are the most important figures for tertiary courses. Their roles include, but are not limited to, designer of the course, presenter of course content, leader of the learning and manager of the class (Dennen, Aubteen, & Smith, 2007). They shoulder responsibility for establishing and maintaining a productive and stimulating learning environment (Dixon, Andreasen, & Stephan, 2009), and their conceptions and practices play a pivotal role in the construction and implementation of the course.

Semi-structured interviews were designed and carried out with LTAC instructors to elicit how they construct and implement LTACs. Compared with structured interviews, which follow a set of rigorous procedures and questions, semi-structured interviews are more open and flexible, allowing interviewees to speak as much as they wish, and are rich data sources for analysis (Edward, 2013). Semi-structured interviews with LTAC instructors were used because: 1) They are flexible and capable of gathering various types of information ranging from factual data, opinions, beliefs, stories, to personal emotions; 2) they engage the researcher and the participants interactively to gather rich data that cannot be collected through focus group discussion and observation (Atkins & Wallace, 2012).

The development and implementation of the semi-structured interview used in this study followed May’s (1991) guidelines. Interview questions were framed to elicit information about how instructors design, construct, and deliver the LTACs and how they cultivate student teachers’ assessment literacy; they included questions about policy and the external cultural and educational environment. Interview questions used simple, comprehensible language related to participants’ experiences. The first draft of the interview protocol was designed by the researcher and reviewed by two professional educational assessment experts and one LTAC instructor for comments and advice. After the revision, the second draft was used as the final interview protocol (see Appendix B).

The interviews were conducted in the instructors' first language, either Chinese or English, to encourage the interviewee's most comprehensive and accurate response (Cortazzi, Pilcher, & Jin, 2011). They were also conducted in a natural and relaxing atmosphere to increase the participant's responsiveness. Prompting and probing following the participant's initial responses (Smith, Chen, & Liu, 2008) was used to gain further insights from interviewees. The interviews, which lasted from 30 to 90 minutes, were audio-recorded with participants' permission and transcribed verbatim in the language of the interview.

3.3.3 Questionnaire for students survey.

A web-based on-line questionnaire called Conception, Self-Efficacy and Practice of Assessment (CSEPoA) was developed and employed to examine student teachers' conceptual model of assessment literacy in the Chinese undergraduate context. The reasons to use a survey for this study are 1) the survey study enabled access to large sample of participants within a short period of time (Munn & Drever, 1999); 2) it can obtain data from a sample large and representative enough to make inferences about the population; (3) it can ensure attention is paid to all the theoretically derived constructs and issues, and (4) reduce effects created by presence of researcher.

3.3.3.1 The development of CSEPoA.

The CSEPoA is composed of the following sections: 1) Demographic information section (6 questions); 2) Teachers' Conceptions of Assessment in Chinese Context (C-TCoA) instrument (32 items); 3) Self-efficacy of Assessment instrument (SEoA) (8 items); 4) Practice of Assessment instrument (PoA) (25 items). The total 64 instrument items (excluding the demographic section questions) are on a 6-point positively-packed rating scale from strongly disagree to strongly agree (Note: the scales are 1=strongly disagree, 2=mostly disagree, 3=slightly agree, 4=moderately agree, 5=mostly agree and 6=strongly agree). This unbalanced positively-packed questionnaire can increase the variation and discrimination of participants' responses (Lam & Klockars, 1982) and avoid bias in the results, especially when participants tend to have an inclination towards positive attitudes (Brown, 2004).

The C-TCoA instrument was developed in Southern China and Hong Kong by Brown and his colleagues (Brown et al., 2011) and has been tested with both Chinese in-service and pre-service teachers (Chen & Brown, 2016). More information about the history of this instrument can be found in section 2.3.3.2, Chapter 2. The SEoA instrument contains only one construct with eight items. The eight items were selected from the Classroom Assessment Confidence Index (CACI) questionnaire (Gu, 2016), which has been used and validated with 239 Chinese primary and secondary teachers. The items were slightly revised based on the findings of Studies 1 and 2. Based on Study 1 and Study 2 findings, it was found that the external testing environment and exam preparation exerts great influence on the course construction and implementation, resulting in a tension between testing and assessment. Thus, PoA took the three factors into consideration and developed corresponding items to measure PSTs' assessment practices in the following constructs: 1) Testing practice; 2) formative assessment practice; and 3) examination preparation practice. The items in PoA were adopted from a number of previously published and validated questionnaires (Brown et al., 2009; Jian & Luo, 2014; McMillan, 2001; McMillan, Myran & Workman (2002); Smith, Hill, Cowie & Gilmore, 2014). Table 2 presents the sources for CSEPoA items in detail. Table 11 in section 6.2, Chapter 6 lists the factors and their definitions.

Table 2

CSEPoA Item and Source

Instrument		Source
C-TCoA		All items adapted from C-TCoA (Brown et al. 2011)
SEoA		All items adapted from CACI (Gu, 2016)
PoA		
Item code	PoA item	PoA item sources
Testing practices		
TP1	I prefer getting tests from an external source rather than writing them myself.	Adapted from McMillan (2001) and McMillan et al. (2002)
TP2	I use tests as a major tool to motivate students to learn.	
TP3	I use tests designed primarily by myself.	
TP4	I value students' test results more than their classroom performance.	
TP5	I look at test results to reflect on and improve my teaching.	
TP6	The major form of feedback I give students is test results.	
TP7	I depend on test results to know what students have learned.	
TP8	My tests are aligned with teaching goals.	
Formative practices		
FP1	I point out areas for improvement to students when I give them feedback.	Adapted from Jian & Luo (2014)
FP2	I share learning goals and success criteria with students to help their learning.	
FP3	I encourage and value students' own evaluation of their work.	Adapted from Teacher Belief of Assessment Smith et al. (2014)
FP4	I observe students' classroom performance to adjust my teaching.	
FP5	I use portfolio assessment to record students' learning process.	

FP6	I check that students understand and use my feedback.	
FP7	I emphasize constructive comments over giving scores or grades.	
FP8	I ask questions in class to check students' understanding.	
FP9	I monitor student peer assessments to make sure they give each other useful comments.	
FP10	I assess according to my teaching objectives.	
FP11	I regularly ask students to give feedback on each other's work.	
Exam preparation		
EP1	I assess according to national or provincial standards or norms.	Adapted from Practices of Assessment Inventory (Brown et al., 2009)
EP2	I try my best to help students get high test scores.	
EP3	The priority of my assessment activities is to help students get good results in their examinations.	
EP4	I assess according to gaokao/zhongkao rules.	
EP5	I use classroom assessment to help student prepare for tests.	
EP6	I regularly use practice tests with students to help them prepare for final exams.	

3.3.3.2 The translation of CSEPoA.

Both English and Mandarin Chinese versions were provided to ensure that PSTs can understand the questionnaire questions well. Both C-TCoA and SEoA have been translated and validated in previous studies (Brown et al., 2011) with both English and Mandarin Chinese versions while most items of PoA were initially written in English. The author translated PoA items from English to Mandarin Chinese and meanwhile, refined the Mandarin translation of C-TCoA and SEoA to make it clearer and more concise. Afterwards, the author asked two professional translators (English to Mandarin Chinese) and a native Mandarin-speaking research colleague who shared the same research interest to check the entire translation and provide as detailed comments as possible. Minor differences in word choice and typographic errors were revised accordingly.

3.3.3.3 The piloting and validation of CSEPoA.

Merten (2015) proposed the questionnaire piloting and validation process should seek both experts' insights and target participants' opinions to: 1) Ensure that the items were well aligned with what was intended; and 2) check the wording and format is appropriate. Based on these principles, the pilot survey was distributed to two LTAC instructors, two educational assessment experts and 30 undergraduate students in year 3 and 4 in one Chinese normal university to: 1) Check the understandability of questionnaire wording; 2) evaluate the equivalence of the Mandarin and English versions; 3) revise questionnaire items to make them clearer and more understandable; 4) test the user-friendliness of the on-line survey format; and 5) record the response time of participants. Feedback identified areas that needed improvement and corresponding actions were taken which included: 1) correcting translation errors of two items; 2) correcting one typo in Chinese; 3) removing one questionnaire item that was reported by five reviewers that asked repeated questions (namely, AC3: assessment results can be depended on from C-TCoA). The internal consistency of this questionnaire was checked by Cronbach's Alpha values and was reported in Chapter 6.3, Chapter 6.

3.4 Data Collection

Data collection for this thesis was conducted in two phases through the following methods and process.

3.4.1 Phase-I qualitative data collection.

The major data underpinning Phase-I data collection included LTAC syllabi, and semi-structured interviews with LTAC instructors. The course syllabi and interviews with instructors were gathered together between November 2016 and March 2017.

3.4.1.1 Purposive sampling and the process.

Purposive sampling is

> A form of non-probability sampling in which decisions concerning the individuals to be included in the sample are taken by the researcher, based upon a variety of criteria which may include specialist knowledge of the research issue, or the capacity and willingness to participate in the research (Jupp, 2006, p.244).

As a widely used technique, purposive sampling is regarded as the most effective method for identifying and selecting the most information-rich cases among the complex phenomena (Etikan, Musa & Alkassim, 2016; Tongco, 2007). It was used to recruit Phase-I participants for two reasons: 1) In the Chinese context, there is theoretically a large number of LTAC instructors, but no recognized data base from which to sample systematically; 2) in the New Zealand context, there are very few LTACs for ESOL teacher education as such courses are not offered by most tertiary institutions. Thus, purposive selection was a feasible and effective method to recruit participants.

The researcher started the collection process by examining official websites of China's and NZ's tertiary institutions to locate ESOL teacher education programmes where LTACs were offered. It was found that, because of the high demand for ESOL teachers in educational sectors, from primary to tertiary, in China, LTACs were offered in both undergraduate and postgraduate level teacher education programmes. In contrast, LTACs were offered at only postgraduate level as part of the post-graduate diploma in Teaching English to Speakers of Other Language (TESOL) or Master of TESOL programmes in NZ.

In China's context, considering the large number of tertiary institutes, the search was confined to the eight state-owned national and 22 provincial normal level universities where professional ESOL teachers were trained. In the New Zealand context, tertiary education is divided into private training establishments (PTEs), institutes of technology and polytechnics (ITPs), universities and wananga (Tertiary institutes incorporating Maori components). The research focused on public-funded universities and polytechnics, which are more comparable to China's national and provincial universities. Meanwhile, the researcher joined national and international academic meetings on language testing and assessment in China and New Zealand to promote this research project and to make connections with LTAC instructors.

The strategies outlined above worked well to attract the attention of interested participants. Invitation letters, Participant Information Sheets and Consent Forms (see appendices A) were sent to potential participants' emails so that they could make an informed decision; interviews were arranged if they kindly agreed to participate. Course syllabi were collected from either the course instructors or downloaded from the relevant institute's websites by the researcher.

3.4.1.2 Study 1 and 2 participants.

Altogether, 20 interviews were conducted with LTAC instructors including 16 from China (ten at undergraduate level and six at postgraduate level) and four from New Zealand postgraduate level; 20 course syllabi from these same course instructors were collected. For confidentiality and anonymity reasons, the instructor participants were labelled with a combination of letters representing:1) Nationalities (C for Chinese; NZ for New Zealand); 2) academic levels (U for undergraduate; P for postgraduate); and 3) numbers indicating the chronological order of the interviews conducted in each nation and each academic group. Each syllabus was given a code by adding a letter S (S for syllabus) after the course instructors' given code. For example, CU1 referred to the first Chinese undergraduate course instructor interviewed among the Chinese undergraduate instructors, while CU1S is the syllabus of CU1. CP3 referred to the third Chinese postgraduate course instructor interviewed among the Chinese postgraduate instructors while CP3S is the syllabus of CP3. Likewise, NZP2 referred to second NZ postgraduate course instructor interviewed among NZ postgraduate instructors while NZP2S the syllabus of NZP2.

The 20 LTAC instructors' academic curriculum vitae were downloaded from their university website or were found by using major search engines such as Google or Baidu (the most prominent search engine in China). The LTAC instructors interviewed had extensive teaching and research experience. All had been teaching for more than ten years, and seven had more than 30 years' experience. Among the 16 Chinese instructors interviewed, 14 held doctorates in education, applied linguistics, or related fields, while three of the four NZ instructors held doctorates. Five of the Chinese interviewees and one NZ interviewee were professors. Six Chinese and two NZ instructors were associate professors. Five Chinese and one NZ participant were lecturers. Their research interests spanned a variety of areas from language testing, assessment, second language vocabulary acquisition, and ESOL teacher education almost exclusively within the applied linguistics area. Table 3 summarizes key information about the instructors' teaching experiences, educational backgrounds, academic titles and research areas.

Table 3

LTAC Instructors' Profiles by Country and Academic Course Level

Instructor	Teaching experience (year)	Education background	Academic title	Research area
Chinese Undergraduate				
CU1	10-15	PhD Applied Linguistics	Lecturer	Language testing and test development
CU2	15-20	PhD Applied Linguistics (U.K.)	AP	Language testing and assessment
CU3	15-20	PhD Applied Linguistics	AP	Educational measurement and ESOL students' cognition
CU4	10-15	PhD Applied Linguistics	Lecturer	Language testing
CU5	10-15	PhD Applied Linguistics (U.K.)	Lecturer	Applied linguistics, teacher knowledge and development, language testing and assessment
CU6	15-20	PhD Applied Linguistics	AP	ESOL education in non-tertiary sector, language testing and assessment, reading in ESOL education
CU7	40	PhD Applied Linguistics	Prof	Language testing, test and textbook development
CU8	20-25	Master Applied Linguistics	Lecturer	ESOL teaching
CU9	10-15	Post-doc experiences and PhD in applied-linguistics	AP	Motivation in ESOL learning and learner autonomy, language testing and assessment
CU10	10-15	PhD Applied Linguistics	Lecturer	Language testing
Chinese Postgraduate				
CP1	15-20	PhD Applied Linguistics	AP	Language testing and listening proficiency
CP2	40	PhD Applied Linguistics	Prof	Language testing
CP3	40	PhD Applied Linguistics	Prof	Language testing and assessment, ESOL teacher education
CP4	40	PhD Applied Linguistics	Prof	Language testing and test development
CP5	20-25	PhD Applied Linguistics	AP	Motivation in ESOL learning and learner autonomy
CP6	20-25	PhD Applied Linguistics	Prof	Language testing and assessment, ESOL teacher education, ESOL classroom research
NZ Postgraduate				
NZP1	> 30	PhD Applied Linguistics	AP	Learner autonomy and learning strategies, language testing and assessment, vocabulary acquisition.
NZP2	> 30	Master Applied Linguistics	Lecturer	Language testing and assessment
NZP3	> 40	PhD Applied Linguistics	Prof	Language testing and assessment, vocabulary assessment, testing for academic and professional purposes
NZP4	> 30	PhD Applied Linguistics	AP	Learner autonomy and learning strategies, language testing and assessment, vocabulary acquisition.

3.4.2 Phase-II quantitative data collection.

According to Creswell and Clark (2017) there are two things to consider for quantitative data collection in exploratory sequential mixed methods research: 1) The participants sampled should be representative to provide information and insights about the phenomenon studied 2) The sample size should be sufficient so the researcher can run statistical analyses to confirm or reject hypotheses. Quantitative data collection normally expects random selection of participants to "remove the potential influence of external variables and ensure generalizability of results" (Creswell, 2009, p.15), however, under constraints of voluntary participation and anonymity, participants' self-selection creates a convenience sample (Suen, Huang, & Lee, 2014). In convenience sampling, large voluntary sampling datasets are commonplace (Etikan et al., 2016). Although this introduces self-selection bias, some control for this can be managed by evaluating the obtained sample characteristics (e.g., age, sex) relative to the population of interest (Hirschfeld & Brown, 2009). Furthermore, large voluntary samples create a robust basis for model estimation in CFA and SEM statistical methods (Neugebauer & Wittes, 1994).

The LTACs are included in the final two years of China's teacher education programmes and, as a result, the target population is senior pre-service ESOL teachers (Year 3 and 4 students and hereafter called PSTs) studying in Bachelor of TESOL programmes (BTESOL) in Chinese normal universities and foreign language studies universities. PSTs were chosen for the research focus in Study 3 because: 1) There are a large number of PSTs in the Chinese undergraduate tertiary sector to meet the large sample size requirement for quantitative survey research so as to better address the address the research questions; and 2) the homogeneity of the research sample requirement, namely, PSTs share quite a lot in common as they had no previous teaching experience or similar learning experiences.

3.4.2.1 Survey channel.

An on-line survey was implemented because of its low-cost and high efficiency compared with paper-based approaches (Fleming & Bowden, 2009), and for its ability to generate a very large number of responses that can increase representativeness of the obtained sample (De Rada, Ariño, & Blasco, 2016). Web-based survey systems are more likely to ensure data quality by ensuring participants do not skip items.

In this study, as the target sample are Chinese undergraduate students born into the web 2.0 generation and frequent cellphone users (Mei, 2016), a combination of survey platform and social network applications was used for survey data collection. In order to respond to the specific Chinese context, the survey platform Qualtrics was combined with two nationwide popular Web 2.0 social network applications, Wechat and QQ for data collection (Mei & Brown, 2018; Harzing, Reiche, & Pudelko, 2013). The dissemination of the survey invitation through Wechat and QQ chat groups was both fast and convenient, ensuring that the obtained sample is more likely to represent the population of interest.

3.4.2.2 Data collection process.

The web-based survey campaign was initiated in June, 2017 and finished by the end of January, 2018, covering one semester of one academic year in China. Recruitment of participants was conducted through the following procedures:

- Advertisements about the study were posted on popular public Chinese social media like QQ, Wechat, and Weibo.
- The researcher approached teachers and tutors at the target universities by either email or friends' recommendation. They were provided with PIS and CF (see Appendix C) of this research and asked for their help in spreading the research advertisement to their students' Wechat or QQ groups.
- Student participants were asked to spread further the survey to their friends and classmates.
- Students taking the questionnaire survey would enter a draw to win prizes of portable chargers or U disks worth about RMB50.

3.4.2.3 Study 3 survey participants.

The survey platform and social network application worked well together and attracted more than 3000 visits from 25 institutes across eight provinces in China. 99.16% of the surveys were completed on cell phones, with the rest from survey links, indicating the popular use of cell phones among contemporary tertiary students and the effectiveness of this survey method. Altogether, 1533 responses were received out of 3068 visits with a response rate of 50.06%. More detailed information on participants' demographic information can be found in section 6.1, Chapter 6.

3.5 Data Analysis

The course syllabi and instructors' interviews in phase-I were analysed separately using related but different qualitative methods of content analysis and thematic analysis. The on-line survey data from phase-II were analyzed using statistical methods of CFA and SEM.

3.5.1 Study 1 content analysis of course syllabi.

Study 1 focused on LTAC syllabi to gain an initial picture of their construction and implementation by examining four key course components: general course information, course objectives, course content and assessment plans. Content analysis was applied as the analytical tool.

3.5.1.1 Content analysis.

As a widely used qualitative analysis method with a long history in research, content analysis is regarded as a quantitative way of analyzing qualitative data (Morgan, 1993; Smith, 2000). It combines the intuition of researchers with the quantitative methods of text analysis (Denzin, 2008). According to Hsieh and Shannon (2005), there are three common approaches of content analysis: conventional, directed, and summative content analysis.

Within conventional content analysis, theories or previous research, are considered in the analysis and discussion process, as the discussion of current study should make a contribution in the same area of research interest. However, there are two challenges: 1) It may fail to generate a complete understanding of the context, or an accurate interpretation of the data; 2) It can be confused with grounded theory method which uses similar initial analytical methods but goes further by looking at the context for nuanced understanding of the lived experience.

Summative content analysis is free of theories and explores the usage of words, and the underlying context, to infer the meaning of text. It starts by counting the frequency of certain words, or content, in textual materials and then tries to understand the contextual use of words to generate categories, themes, and patterns for discussion.

Directed content analysis is similar to conventional content analysis by relying on existing theories and previous research but is carried out deductively. The researcher identifies key concepts as initial codes and then forms categories defined by the theory applied. Directed content analysis is framed by the theories, and, as a result, the researcher may look for supporting evidence for a particular theory while neglecting non-supporting evidence. In this thesis, Davis's (2008) knowledge + principle + skill AL framework (refer to section 2.3.1.1 Chapter 2 for details) was applied as a theoretical analysis framework, thus the directed content analysis was employed as the analysis method.

3.5.1.2 Content analysis process.

Among the 20 course syllabi, six were written in Chinese, while the rest were in English. I translated the Chinese syllabi to English and sought professional comments from two experienced Chinese-to-English translators for revision and correction. To enhance trustworthiness, I invited one PhD graduate, with similar research interests, as a critical research peer to work together in conducting the analysis. Peer debriefing, prolonged engagement, and checking were carried out at each step to minimize individual bias and errors.

During the analysis, as one principle is to look at how words and meanings are used in these contexts and to draw inferences of the underlying focus and patterns with a comparative lens, the focus was on the message content rather than the format (Johnson & Weller, 2002). Following the evaluation framework for course syllabi created for this study (see Figure 2 in section 3.3.1), texts from each component (general information, course objectives, course content and assessment plans) were coded into explicit items. This initial work laid the foundation for generating subsequent categories and patterns. After the initial coding, key words were written down as they appeared in the syllabi. The on-going list of key words was counted to reveal the prevalence of each topic and put into groups and categories based on the evaluation framework. Finally, similar items were grouped together in the same category, with the grouping process continuing until there was no further new information. The findings are reported in Chapter 4.

3.5.2 Study 2 thematic analysis of instructors' interviews.

The Study 2 interview data were subjected to thematic analysis. As for the translation of the course syllabi, I translated the key points of the Chinese instructors' interview transcripts into English and sought professional comments from two experienced Chinese-to-English translators for revision and correction.

Tesch (1990) described two traditions of qualitative analysis, the linguistic tradition in which texts are regarded as an object of analysis and the sociological tradition in which texts are a window into human experience. An inductive thematic analysis (Braun & Clarke, 2006; 2013) using the second approach, the sociological tradition, was employed to seek the themes and stories embedded in the instructors' interviews for their thoughts and experiences.

Braun & Clarke (2006) argued that thematic analysis is an essential method for qualitative research because it contains the core skills to conduct analysis of various types of qualitative data. Thematic analysis is highly flexible for it does not require detailed theoretical and technological considerations as do other qualitative approaches. It can be modified to be applicable to a wide range of research paradigms and research questions; and a strong thematic analysis can provide highly reliable and insightful findings of complicated data (Braun & Clarke, 2013). Inductive thematic analysis was employed in this study for its flexibility and compatibility to fit a variety of frameworks, questions, people and phenomena (Braun & Clarke, 2006). Themes are the abstract constructs that the researcher can identify before, during and after data collection and are formed or identified inductively from the text data.

To ensure trustworthiness, a six-phase analysis process including data familiarization, initial code generation, theme search, theme review, theme definition and naming and theme writing up was followed by constant checking and reflecting in between each step. A research colleague, fluent both in Chinese and English with research interest in language assessment, and one assessment research expert (the supervisor) were closely involved in the analysis process. Discussions with the research colleague and the assessment expert were constantly conducted to verify accuracy and authenticity of the interpretation of the data. Table 4 presents in detail how each phase was carried out for the thematic analysis process. The findings are reported in Chapter 5.

Table 4

Thematic Analysis Process in Study 2

Phase	Phases	Details to ensure trustworthiness
1	familiarizing oneself with the data	Read each transcript at least 5 times Store raw data in well-organized archives Write down reflective thoughts Document thoughts about potential codes/themes Research colleague reads two transcripts and makes notes
2	generating initial coding	Use an open coding method Audit trail of code generation Review initial coding with research colleague Discuss initial coding with research colleague
3	searching for themes	Put codes into categories and then themes Keep detailed notes about development and hierarchies of concepts and themes Discuss with research colleague
4	reviewing themes	Themes and subthemes verified by research colleague Test for referential adequacy by returning to raw data
5	defining and renaming themes	Reach consensus on themes with colleague Revise the initial definition of themes
6	writing up themes	Describe coding and analysis process in sufficient details Thick description of context developed Colleague and expert (supervisor) checking

Note. Based on the process proposed by Braun and Clarke (2006).

3.5.3 Study 3 quantitative analysis of student survey.

The statistical methods applied for data analysis in Study 3 were confirmatory factor analysis (CFA) and structural equation modelling (SEM). Social science researchers are usually interested in studying the seemingly complex theoretical constructs such as conceptions, beliefs and intelligence which cannot be observed directly. These abstract constructs are called latent or unobserved variables (Byrne, 2016). Latent variables (factors) are not directly observed but are inferred from manifest variables. Manifest or observed variables (items) are directly measured by the researchers with survey items and are used to reflect the constructs. Exogenous variables are independent variables presumed to cause changes in dependent or endogenous variables. In contrast, endogenous variables are dependent, whose values are determined by others (Byrne, 2016). Exogenous variables may be caused by other factors, but such causes cannot be explicitly modeled from the data available.

The relationship between observed and unobserved variables can be displayed through path diagrams. Manifest variables are usually represented by rectangles or squares while the latent ones are shown as ovals or circles. Residuals, which are also inferred, are represented within ellipses. Curved bidirectional arcs represent correlations, while straight unidirectional lines represent regressions.

3.5.3.1 Confirmatory factor analysis (CFA).

Confirmatory Factor Analysis (CFA) establishes the degree to which a set of items are sufficiently inter-correlated to form a latent factor and how latent factors are inter-correlated. CFA calculation reveals the differences between the estimated covariance matrices with the observed covariance matrices. The confirmed CFA models are called measurement models showing the valid and reliable relationship between items and latent factors (Kline, 2015; Steinmetz et al., 2009). Schreiber et al. (2006) argued that CFA should be theory driven, which means the presumed relationship between observed and unobserved variables should be determined by their theoretical relationships either from theory or from previous studies. The less discrepancy there is between the theory-driven model and the actual observed data matrix, the more likely the model is correct (Kaplan, 2009). CFA can also examine the factor and item validity by assessing both convergent and discriminant validity (Bentler, 1978).

Study 3 intends to test whether the PSTs' responses to conceptions, self-efficacy and practices of assessment fit the previously established measurement model rather than explore new measurement models, thus CFA is employed as it provides information about how well the model fit the data and allows tighter specification of multiple hierarchies or of factor relationships (Klem, 2000; Hoyle, 1995).

Also, inter-correlations between latent factors in each measurement models are calculated in AMOS to explore the relationship between latent factors. According to Taylor (1990), correlation coefficients (*r*) higher or equal to 0.7 are regarded as strong positive relationship, between 0.4 and 0.69 strong positive relationship, between 0.30 and 0.39 moderate positive, between 0.2 and 0.29 weak positive relationship and 0.19 and -0.19 no or negligible relationship; -0.20 and -0.29 weak negative relationship and -0.30 and -0.39 moderate negative relationship.

3.5.3.2 Structural equation modelling (SEM) and invariance testing.

Structural equation modelling (SEM) "expressed either diagrammatically or mathematically via a series of equations" (Byrne, 2016, P.7), describes how various measurement model components relate to each other. SEM is an extension of the general linear model (GLM) but is able to simultaneously estimate regressions, correlations, and residuals among multiple latent and manifest variables (Byrne, 2016). An SEM can be a second-step model after each of the components are reliably identified in CFA measurement models. How various structures interact should be determined first by reference to theory and previous empirical data, rather than just arise from chance artefacts within the obtained dataset (Bandalos & Finney, 2010).

Once a model has been identified, the fit of the model to the data is determined by reference to a series of fit indices.

> It should be noted that a perfect fit hardly exists and thus the differential between the observed sample data and the hypothesized model is what we call the residual. Therefore, the model-fitting process can be summarized as the following equation: Data=Model + Residual (Byrne, 2016, p. 7).

SEM was conducted based on three possible assumptions to establish the structural relationship between conceptions, self-efficacy and practices of assessment (see section 6.4 in detail). To test whether there's equivalence between LTAC-participants and non-participants, multi-group nested invariance testing was conducted based on the confirmed SEM model. Multi-group invariance testing within SEM is the evaluation of models for equivalence between groups, which allows determination of whether observed differences vary within chance or not (Byrne, 2004; Byrne, 2008). It is used to calculate whether a given measure can be treated as equivalent by respondents from different groups (Hirschfeld & Brown, 2009).

Violations of measurement invariance may preclude meaningful comparison of scale scores and relations (Milfont & Fischer, 2010).

3.5.3.3 Model fit indices.

Model fit indices in CFA and SEM analysis are used to determine the invariance and fitness of the hypothesized model and the specific sample data or tested model (Barrett, 2007). Numerous sets of fit indices have been employed by social science researchers and there is no consensus reached on what combination of indices best determine the model fitness (Hooper, Coughlan & Mullen, 2008a). The model fit indices can be categorized as badness-of-fit measure when the decrease of the indices leads to better fit of the model and goodness-of-fit measure when it's the opposite. Badness-of fit indices usually include the chi-square (χ2), Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR) and the goodness-of-fit include Comparative Fit Index (CFI) and Gamma Hat ($\boldsymbol{g}\hat{}$) (Fan & Sivo, 2007). As suggested by Yuan (2005), researchers should use the indices that are appropriate and specific for the individual research data, thus this study chose the following measure as model fit indices.

Chi-square ratio (χ2/*df*) i.e. Chi-square squared divided by degrees of freedom with a probability level at 0.001 is used to estimate the overall model fit. Even though Chi-square can be used as a model fit index, there are severe limitations of its use. First, Chi-square test assumes multivariate normality. Deviation from normality may lead to rejection of the model even when the model is proper (Hooper, Coughlan & Mullen, 2008b). Second, Chi-square statistic is sensitive to sample size and tends to reject the model of large sample size (Bollen & Long, 1993). An alternative to Chi-square used by researchers is the normed chi-square (χ2/*df*) with a recommendation of acceptance range between 3.0 and 3.8 with the *p*-value < 0.001 (Singh, 2009).

The root mean square error of approximation (RMSEA) has become one of the most informative fit indices for its sensitivity to the number of estimated parameters in the model. It tells how well the model with optimally chosen parameter estimates would fit the population's covariance matrix (Byren, 2010). A cut-off value of RMSEA ≤ .08 is regarded as acceptable while ≤ .05 suggests a good fit.

Both root mean square residual (RMR) and standardized root mean residues (SRMR) measures the average value of the standardised differences between observed correlations and the predicted correlations in the model. RMR is based on the scales of the indicators and can be difficult to interpret when there are varying sets of scales (for example, some items range from 1-6 while others from 1-5) (Kline, 2015). As a result, SRMR is usually reported for its stability. SRMR value of 0.08 or less is indicative of an acceptable model and a cut-off value close to 0.06 is the general indicator of good fit (Hu & Bentler, 1999).

Comparative Fit Index (CFI) is used to test how the tested model is better than the alternative model when all observed variables are deemed uncorrelated. A cut-off value of CFI ≥ 0.90 is regarded as acceptable among researchers while 0.95 or above suggests a good fit (Byrne, 2010).

Gamma Hat is related to RMSEA value, degree of freedom and the number of observed variables. It is also stable across model complexity, sample size and model misspecification (Fan & Sivo, 2007). According to Marsh, Wen and Hau (2004), gamma hat ≥ .90 constitutes acceptable fit while gamma hat ≥ .95 a good fit to the data. Based on the above discussion, the model fit indices and criteria for acceptable and good chosen for this study are listed in Table 5.

Table 5

Model Fit Indices Chosen for Study 3

Index	Criteria for good and acceptable
χ2/*df* (*p*<0.001)	< 3 good; < 3.80 acceptable
CFI & Gamma Hat ($\hat{\boldsymbol{g}}$)	≥0.90 acceptable; ≥ .95 good
RMSEA	≤ .08 acceptable; ≤ .05 good
SRMR	≤ .08 acceptable; ≤ .06 good

Note. CFI=comparative fit index; RMSEA=root mean square error of approximation; SRMR= standardized root mean residual; AIC=Akaike information criterion.

3.5.3.4 Model modification.

In model generating, the theoretical specification seldom matches the data in desired fit, thus item trimming can be used to refine survey instruments while maintaining to the greatest extent the underlying hypothesized model (Brown et al, 2017). As Beavers et al. (2013) argued, retaining as many factors as the theoretical model is “less detrimental to the analysis than eliminating factors that are needed” (p. 7). This can help researchers best confirm or reject a model hypothesis without necessarily modifying the conceptual model. When a model solves but does not have a good fit to the data, it is possible to use various techniques to identify a well-fitting model. Poor item fit is identified when 1) items have standardized regression weights less than 0.30 on their conceptual factor (Haertel, Walberg & Haertel, 1991) ; 2) items have large modification indices on other constructs (Bandalos & Finney, 2010); and 4) items have an improper theoretical fit with its factor (Brown et al., 2009). Unfit items were removed one by one to test the model fit indices until a fit model was established.

3.6 Ethical Considerations

Research involving human participants should consider the dignity, health and self-esteem of the participants and it is the researcher’s responsibility to anticipate the possible ethical issues that may be encountered during the research process (Creswell, 2013). To this end and in these contexts, the research design and implementation followed the ethical standards of the University of Auckland to avoid potential dilemmas and, at the same time, put forward possible solutions. Two ethical applications were submitted and approved by Human Participants Ethics Committee (HPEC) at University of Auckland for Phase-I and Phase-II data collection respectively (See Appendix A and Appendix C). During the research, I adhered to the principles required by the ethical standards set by the HPEC, listed as follows:

- Participant Information Sheets and Consent Forms were sent to the course instructors, faculty deans and survey participants’ teachers and tutors to explain the procedure and purpose of this research and asked for their consent and support.

- Participation in this research was voluntary and any participant retained the right to participate or to withdraw. Participants maintained the right to ask questions during the interview and the survey. It has been assured that participation or non-participation had no impact on their course curriculum, grades, employment status or well-being.
- Participants' privacy, confidentiality and anonymity were carefully protected by the researcher. The real names of course instructors were not used and no clues given as to which university or college the course instructors belonged to. Survey participants conducted the web-based survey without writing down their names or any identification information.
- Data were stored and protected in either the researcher's personal password protected computer or her locked office cabinet in New Zealand. Consent forms and data would be stored for six years after completion of the research, and deleted after this time, unless the researcher continues with research in this field. Data collected were used for academic and educational purposes, such as the researcher's PhD thesis at the University of Auckland, academic publications, conference presentations, teaching, and other forms of academic research dissemination.

3.7 Chapter Summary

This chapter started with an explanation of the pragmatism paradigm and mixed-methods research approach used for the thesis, and then introduced the exploratory sequential mixed-methods research design structure. It went on to expound the instrumentation considerations, data collection phases, including the sampling and characteristics of the participants, and the analysis methods used for the three studies; it ended with the ethical considerations for this thesis. With the two phases of data collection process (qualitative and quantitative data), the three studies and different analytical methods, this thesis forms a data triangulation to provide a more comprehensive understanding of the LTACs in different contexts. The following three chapters report findings of Study 1, 2 and 3 to address the overarching and sub research questions.

Chapter 4 Findings of Study 1: Content Analysis of Course Syllabi

This chapter reports the findings of Study 1 in which 20 LTAC syllabi collected from New Zealand and China (ten at Chinese undergraduate level, six at Chinese postgraduate level and four at NZ postgraduate level) were analysed using content analysis. Based on the evaluation framework developed for Study 1 (refer to section 3.3.1, Chapter 3), patterns and features among syllabi were examined to identify similarities and differences in course design and in how they worked to develop students' assessment literacy. The overarching research question asked: "How are LTACs constructed and implemented in different cultural and academic contexts and how do they shape students' assessment literacy?" and thus the sub-questions for this study were proposed as follows:

1. What are the key messages and patterns conveyed in the core components (namely, general course information, course objectives, contents, and assessment plans) of LTAC syllabi?
2. How might these aspects of syllabi contribute to the development of student teachers' assessment literacy?
3. What are the similarities and differences between LTACs in different cultural and academic contexts as reflected in the syllabi?

The following sections report the analysis results of each core course component across the three contexts of courses: Chinese undergraduate courses (CUCs) are presented first, followed by the Chinese postgraduate courses (CPCs), and lastly the New Zealand postgraduate courses (NZPCs). The final part in each section summarizes and compares the findings with comments.

4.1 General Course Information

This section provides a general overview of LTACs in each context with information on course title, timeframe (delivery time and course length), course status (compulsory or optional) and expected workload for this course.

4.1.1 Chinese undergraduate course information.

The ten CUCs were a component of the specialized courses set for the Bachelor of TESOL education programmes (BTESOL) in the participating Chinese universities[2]. The word "testing" was found commonly in the course titles across the CUCs. Seven out of ten courses were called "*Language Testing*" while three were entitled "*Language testing and assessment*". Eight courses were compulsory whereas two were optional. The undergraduate courses lasted between eight to 16 weeks during one semester and were scheduled either before or after student practicum during the final two years of the four-year teacher education programmes. Only four of the ten CUCs specified the expected learning hours, which were between 16 hours and 36 hours of class contact time.

4.1.2 Chinese postgraduate course information.

The six CPCs were components of the Master of Applied Linguistics or Master of TESOL education programmes, which were between two to three years full time in these Chinese universities. There was a greater variety of course titles in CPCs than in the CUCs. Of the six CPCs, three were titled "Language testing and assessment"; two "Language testing" and one "Assessment measurement". CP1S (course syllabi of the first instructor interviewed in the Chinese postgraduate courses[3]) was co-designed with a U.K. university and delivered by Chinese instructors on a campus in China. Four of the six CPCs were compulsory. All the six CPUs were delivered on-site rather than on-line, lasting one semester between eight to ten weeks. Two of six CPCs specified the expected workload: CP1S required 24 hours of in-class learning and CP4 36 hours.

[2] In BTESOL, students take English language courses such as extensive English and intensive English to enhance language skills in the first two years; later they continue with specialized courses such as English language pedagogy and language testing and assessment that prepare them for future teaching.

[3] Refer to 3.4.1.2 for details regarding labelling of participants and course syllabi.

4.1.3 NZ postgraduate course information.

Similar to the CPCs, the four NZPCs were papers required in the Master of Applied Linguistics, Master of TESOL and Postgraduate Diploma of TESOL programmes in the participating New Zealand universities and polytechnic. There was a wide variety of course titles. Each NZ course had a different title: “*Language assessment*”, “*Language policy and assessment*”, *“Language testing” and “Language assessment*” (one NZ instructor had two versions of LTACs) and “*Evaluating and developing language assessment tasks*”. All NZPCs were compulsory for students taking the programmes. Students could choose to complete the entire programme in two years full time or four years part time, and had freedom to choose when to attend the course. The NZPCs were partially or entirely delivered on-line, which were implemented and scaffolded by the unique on-line learning environment in each institution. The courses were from one trimester to one semester (either 10 or 16 weeks) in New Zealand. All NZPCs specified that study hours were made up of in-class learning, on-line learning and after-class independent learning. The overall expected learning hours were between 150 to 300 hours during which independent learning occupied around half of the time.

4.1.4 Summary and comments.

In summary, most CUCs were called “language testing” whereas CPCs were mostly called “language testing and assessment”. New Zealand course titles had the greatest diversity among the courses. In terms of course length, there was little difference between countries or levels, with most courses taking between 8 and 12 weeks. Similarly, across countries and levels most courses were compulsory rather than optional. Most courses in China (at both undergraduate and postgraduate levels) did not specify expected workload or learning hours. In contrast, NZ course syllabi specified learning hours between 150-300 hours and required a large amount of independent learning hours. NZPCs emphasized the importance of after-class independent learning. All NZPCs incorporated on-line platforms to facilitate teaching and learning while this was not the case in the Chinese courses (at both undergraduate and postgraduate level). Table 6 summarizes the general course information across all syllabi.

Table 6

General Course Information by Country and Academic Level

Course code*	Course title	Course status	Timeframe	Expected workload
Chinese undergraduate course syllabi				
CU1S	*Language testing and assessment*	Compulsory	Semester 2, Year 4 (10 week)	Not specified
CU2S	*Language testing*	Compulsory	Semester 1, Year 4(10 weeks)	Not specified
CU3S	*Language testing*	Compulsory	Semester 1, Year 3 (12 weeks)	36 hours in-class learning
CU4S	*Language testing*	Compulsory	Semester 1, Year 4 (8 weeks)	16 hours in-class learning
CU5S	*Language testing*	Optional	Semester 1, Year 3 (12 weeks)	Not specified
CU6S	*Language testing and assessment*	Optional	Semester 1, Year 3 (8 weeks)	Not specified
CU7S	*Language testing*	Compulsory	Semester 1, Year 3 (10 weeks)	Not specified
CU8S	*Language testing*	Compulsory	Semester 1, Year 4 (10 weeks)	26 hours in-class learning
CU9S	*Language testing and assessment*	Compulsory	Semester 2, Year 4 (12 weeks)	Not specified
CU10S	*Language testing*	Compulsory	Semester 2, Year 3 (17 weeks)	32 hours in-class learning
Chinese postgraduate course syllabi				
CP1S	*Assessment and measurement*	Compulsory	Semester 2 (12 weeks)	24 hours of in-class learning
CP2S	*Language testing and assessment*	Optional	Semester 2 (8 weeks)	Not specified
CP3S	*Language testing and assessment*	Optional	Semester 2 (16 weeks)	Not specified
CP4S	*Language testing*	Compulsory	Semester 2 12 weeks	36 hours of learning

Course code*	Course title	Course status	Timeframe	Expected workload
CP5S	*Language testing and assessment*	Compulsory	Semester 1 (12 weeks)	Not specified
CP6S	*Language testing and assessment*	Compulsory	Semester 1 (10 weeks)	Not specified
NZ postgraduate course syllabi				
NZP1S	*Language testing; Language assessment*	Compulsory	One trimester (10 weeks)	150 learning hours: 30 on class contact or participating in on-line learning, 50 on reading and 70 on assignments
NZP2S	*Evaluating and developing language assessment tasks*	Compulsory	One semester (12 weeks)	300 learning hours: 45 on class contact; 45 on-line learning and 210 self-study
NZP3S	*Language Policy and assessment*	Compulsory	One semester (16 weeks)	300 learning hours with on-line learning component
NZP4S	*Language assessment*	Compulsory	One semester (12 weeks)	150 learning hours

**Note.* Refer to section 3.4.1.2, Chapter 3 for course code labeling.

4.2 Course Objectives

In education programmes, teaching objectives and learning outcomes describe the expectations of courses, but from different perspectives. Teaching objectives are proposed from the instructors' perspectives in terms of what they intend to teach while the learning objectives refer to what students should achieve by the end of the course. In the collected syllabi, course objectives were named in several different ways such as "teaching objectives", "learning outcomes" and "learning goals" but they all described the ultimate goals and purposes that direct teaching and learning.

An overall 26 types of course objectives were summarized through the analysis across the 20 course syllabi; seven were about knowledge, nine about principles and ten about skills. Objectives in the knowledge dimension refer to knowing the fundamental knowledge in language measurement and description such as the concepts, types and purposes. Objectives in the principle dimension refer to understanding the proper use of assessment including the validation principles, the relationship between tests, assessments and teaching, the social dimension of assessment and the ethics questions and so on. Objectives in the skill dimension refer to the ability to carry out assessment-related tasks such as test or assessment development, analysis, interpretation and communication. Table 7 presents the types of objectives and their frequency (the times certain objectives appeared across the syllabi) in terms of the dimensions of knowledge, principle and skill in CUCs, CPCs, and NZPCs.

Table 7

Course Objectives by Country and Academic Level

Objective	Frequency		
	CUC (n=10)	CPC (n=6)	NZPC (n=4)
The knowledge (aim to learn…)			
1. (Basic)* fundamental concepts of language testing	8	1	
2. Different type of tests and its purposes and functions	4	3	
3. Test design process	2		1
4. (Basic) theory of language testing research	1	1	
5. (Basic) knowledge of language testing and assessment	2	3	
6. Various methods to language assessment		1	3
7. Current trends and issues in language assessment			2
The principle (aim to understand…)			
1. (Basic) theory in language testing	4		
2. Relationship between testing and teaching	5	3	
3. Principles for test validation	3	2	1
4. Relationship between learning, assessment and teaching		2	
5. Relationship between language assessment and policy			1
6. Relationship between curriculum and language assessment			2
7. Language assessment development principles		3	1
8. The social and educational impact of language assessment			1
9. Ethics in language assessment		2	1
The skill (to be able to…)			
1. Design tests for teaching	7	4	2
2. Analyse test quality (validity, reliability, authenticity, etc.)	5	3	2
3. Interpret test results	5	2	
4. Compute statistics for language testing analysis	2	2	1
5. Design assessment tasks for teaching	2	3	3
6. Use multiple ways of assessment		2	1
7. Critically evaluate assessment tasks or practices			2
8. Conduct research in language testing and assessment		3	2
9. Link assessment practice with theoretical issues			1
10. Assess native and non-native speakers			1

Note. (Basic) only appeared in the undergraduate level.

The course objectives across all syllabi were either about tests/testing (hereafter testing) or assessments/assessing (hereafter assessment). After comparing the frequency of testing and assessment-related objectives in each context, a pattern of different foci emerged from the analysis. Within the CUCs, there was a dominant percentage of objectives in testing (92% = 46/50) while only 8% objectives were about assessment (4/50). Within the CPCs, there were slightly more testing-related objectives (58.7% = 27/46) than assessment-related (41.3% = 19/46). NZPCs had the greatest amount of assessment-related objectives (75% = 21/28) among the courses while only 25% (7/28) were about testing. Figure 3 illustrates the foci across the courses in different contexts.

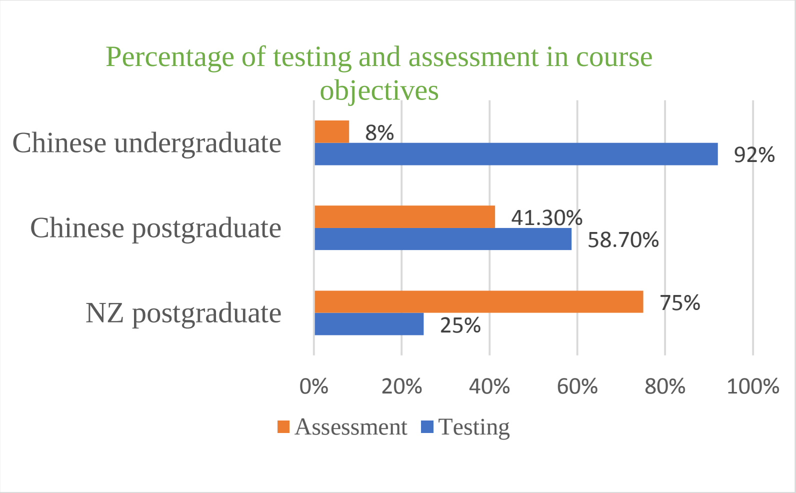

Figure 3. Percentage of testing and assessment in course objectives across syllabi.

Calculation of the frequency of objectives in the three dimensions of knowledge, principles and skills across all syllabi showed that the largest proportion of objectives were in skill (46.6% = 55/118), followed by knowledge (27.1% = 32/118) and principles (26.3% = 31/118). See Figure 4 for details.

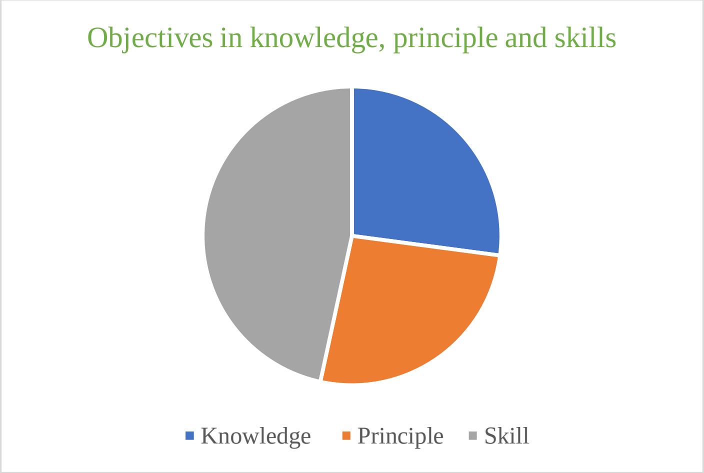

Figure 4. Proportion of objectives in three dimensions.

4.2.1 Chinese undergraduate course objectives.

The ten CUCs had a dominant focus on testing-related objectives, with an emphasis on the practical application of skills compared with the learning of knowledge and principles. The objectives were at a basic level as reflected in the wording, for example, “basic concepts”, “basic knowledge” and “basic theory”.

Within CUCs, there were 13 types of course objectives across knowledge, principles and skills of which nine were about testing and four were related to assessment. The objective with the greatest frequency was to understand the basic and fundamental concepts of language testing (8 /10 courses); the second related to designing tests for teaching (7/10), while the third was to understand the relationship between testing and teaching (5/10). In the knowledge dimension, basic knowledge of testing was most-frequently mentioned with priority (8/10), followed by the learning of different types and functions of tests (4/10). Two out of ten course syllabi emphasized knowing the process of test design and one syllabus required students to acquire knowledge about language testing research. Only two out of ten syllabi mentioned the word assessment, stating that they required students to learn knowledge of language testing and assessment rather than focus only on language testing. In the principle dimension, the most frequent objective concerned understanding the relationship between testing and teaching (5/10), followed by the basic theory of language testing (4/10) and the principles of test validation (3/10). In the skill dimension, the objectives were to prepare teachers’ language testing skills mainly in three ways: 1) To design basic tests for teaching (7/10); 2) to analyse and evaluate test quality (5/10); and 3) to interpret test results (5/10). Two of these courses included basic statistical analysis ability as learning objectives to enhance the implementation of these skills. Only two of the ten courses set designing assessment tasks for teaching as objectives.

Calculation of frequency of objectives within the CUCs revealed a similar pattern with the overall proportion across all syllabi as shown in *Figure 4*. There was a greater percentage of objectives related to skills than either knowledge or principle in CUCs. As the percentage shows, 42% (21/50) of the objectives were categorized in the skill dimension; 34% in knowledge (17/50) and 24% (12/50) in principle.

4.2.2 Chinese postgraduate course objectives.

Throughout the six CPCs, there was a greater emphasis on assessment, a greater variety of objectives, and a similar focus on objectives of skills in comparison with CUCs. Across the 17 types of objectives within CPCs, more than half (9/17) were assessment-related. Assessment was mentioned, together with testing, to indicate the importance of non-testing assessment methods. Half the syllabi required students to: 1) possess knowledge of language testing and assessment; 2) understand language assessment development principles; 3) design assessment tasks for teaching; and 4) conduct research in language testing and assessment. One third of the syllabi asked students to: 1) understand the relationship between learning, assessment and teaching; 2) understand ethics in language assessment; and 2) use multiple ways of assessment. The emphasis on assessment was evident in quotes from the syllabi: “This module aims to provide an understanding of a range of assessment practices and techniques” (CP1S); “to use multiple ways of assessment to support and assess students’ learning (CP6S)”. There was also a greater variety of objectives in the postgraduate course syllabi.

There was a consistent emphasis on assessment in the knowledge and principle dimensions as well, with assessment related objectives mentioned in the skill dimension. In the knowledge dimension, knowledge of language testing and assessment was noted most frequently (3/6) along with the types and purposes with tests (3/6). In the principle dimension, similar to the CUCs, the relationship between testing and teaching was emphasized with the greatest frequency (3/6). In contrast to CUCs, CPCs took assessment into consideration with two out of six courses requiring an understanding of relationship between learning, assessment and teaching. More than half (4/6) courses required students to design tests for teaching; half of the courses (3/6) demanded skills to analyse test quality and one third (2/6) required the skills to interpret test results and to compute statistics for test analysis. Ethics in language assessment was first raised as an objective for CPCs (2/6). Half of the syllabi required students to design assessment tasks while one third referred to using multiple ways of assessment. In comparison with CUCs research was given greater importance; half the course syllabi required students to conduct research on language testing and assessment.

In summary, the frequency of the course objectives in CPCs across the three dimensions of knowledge, principles and skills revealed a similar focus as appeared in CUCs. The emphasis was on skill rather than knowledge and skills: 47.5% (19/40) of objectives were concerning skills; 22.5% (9/40) knowledge and 30% (12/40) principles.

4.2.3 NZ postgraduate course objectives.

There was a dominance of assessment-related objectives in NZPCs and a greater variety of objectives compared with CUCs and CPCs. Although there were only four course syllabi in this group, 18 types of objectives were identified among which 14 were about assessment.

In NZPCs, the word assessment was used throughout the syllabi. The most frequently mentioned course objectives were to 1) understand various approaches to language assessment (3/4) and 2) design assessment tasks for teaching (3/4). An explanation can be found in quotes from NZP2S: "This module requires participants to develop, implement, evaluate and modify assessment tasks suitable for assessing language skills." NZPCs had the broadest range of objectives among the three groups of courses, ranging from social and educational impact of language assessment to the relationship between language assessment and its mediating factors like policy and curriculum.

In the knowledge dimension, the emphasis was on knowing various methods of language assessment (3/4) and the current trends/issues in language assessment (2/4). In the principles dimension, NZPCs talked about the wider social educational factors that interacted with language assessment, namely, the relationship and interaction between language education, policy, and curriculum. Half of the syllabi (2/4) put the relationship between curriculum standards and language assessment as a course objective, and one required students to know the relationship between language assessment and policy. For example, NZP1S required students to "understand how language assessment relates to curriculum objectives in language education." NZP3S required students to understand "the inter-relationship between language policy and language assessment", and to "reflect on theory, issues, practices and current trends in language assessment." NZP4S asked students to "link assessment practice with theoretical issues" and "to be familiar with key literature in the field".

In the skill dimension, there were requirements for designing assessment tasks for teaching (3/4) as well as analysing and evaluating assessment tasks (2/4). There were also requirements for language testing: For example, half of the syllabi (2/4) specified that students needed to be able to 1) design tests for teaching; and 2) analyse test quality. Language assessment theory was also mentioned. Context and purpose, however, were emphasized as well as in these objectives. As cited in the course syllabi, the courses aimed: "To design "high-quality tests for particular purposes" (NZP1S); to critically evaluate language tests and their suitability for particular contexts (NZP3S); and to "demonstrate an ability to plan appropriate assessment for a learning context you are familiar with (NZP2S)". Research was also a focus in NZPCs with half of the course syllabi requiring students to conduct research in this field.

NZPCs had a similar focus on skills of practical application as the CUCs and CPCs. The frequency of the course objectives in this group was: 53.6% (15/28) of the objectives were categorized in the skill dimension; 21.4% (6/28) in knowledge and 25% (7/28) in principle.

4.2.4 Summary and comments.

To conclude, at the Chinese undergraduate level, course objectives had a consistent focus on language testing, whereas there were more objectives concerning the wider conceptions of assessment at Chinese postgraduate level. In NZ postgraduate level, there appeared to be a broader range of course objectives using language assessment as the umbrella term that incorporated testing and other forms of assessment approaches and issues. This suggested that NZPCs used language assessment as a tool for assessing students' language, regardless of its form. Inspection of objectives across the three dimensions of assessment revealed there was an emphasis on assessment skills, compared with the assessment knowledge and principles, ranging from test/assessment design, test/assessment analysis/evaluation to test/assessment results interpretation.

Comparison between the two countries identified the following differences: 1) NZPCs used the term assessment more often as an umbrella term incorporating testing and assessment, whereas Chinese courses (including the Chinese undergraduate and postgraduate groups) used testing and assessment separately; and 2) NZPCs required students to understand broad and diversified issues such as the social and educational impact of language assessment, relationship between teaching, assessment and contextual factors like policy and curriculum, whereas in the Chinese context students were expected to understand the relationship between language teaching and testing or assessment. This expectation reflects the influential washback effect of testing in China.

Comparing the two academic levels, there was a greater variety of objectives in postgraduate courses (Chinese postgraduate and NZ postgraduate) than in the undergraduate courses. There were 13 types of objectives in CUCs, whereas 17 in CPCs and 18 in NZPCs respectively. Furthermore, there were more assessment-related objectives in the postgraduate level courses. Postgraduate courses also focused more on research and theory, requiring higher-order thinking skills whereas the undergraduate courses set the "basic" requirement of language testing and assessment.

4.3 Course Content

Course content reflects instructors’ choice of topics to support course objectives. Based on the evaluation framework for Study 1, the analysis identified 36 types of content topics, which were summarized and categorized 14 in knowledge, nine in principle and 13 in skill. These content topics covered the fundamental knowledge, principles and skills students need to develop under language assessment literacy and were aligned with course objectives across the three groups of courses. Table 8 presents the types of content topics and their frequency in each context of courses (CUCs, CPCs, and NZPCs).

Table 8

Course Content by Country and Academic Level

Content topic	Frequency		
	CUC (n=10)	CPC (n=6)	NZPC (n=4)
Knowledge			
1. Concepts of language testing	9	4	
2. Purposes and functions of tests	7	2	
3. Purposes of assessments			3
4. Types of tests	8	2	
5. Types of assessments			3
6. History of testing	3	2	
7. Types of test questions	3	1	
8. Test specification	6	1	1
9. Test design process	5	2	1
10. Formative assessment	3	4	2
11. Classroom assessment	1	1	1
12. Portfolio as assessment	2	2	2
13. Computerized assessment		1	1
14. Current trends and issues in LTA		3	2
Principles			
1. Test validity & reliability	5	2	2
2. Features of a good test	2	1	1
3. Test validation principles	3	2	2
4. Relationship between language testing and	6	3	
5. Relationship between assessment and			2
6. Assessment and policy/curriculum			2
7. Language assessment principles	1	3	2
8. The social impact of assessment			2
9. The ethics considerations in LTA			2
Skills			
1. Testing listening, speaking, reading, writing	6	3	1
2. Testing grammar and vocabulary	4	1	1
3. Item writing and analysis	5	2	
4. Statistical analysis	3	3	2
5. Tests results interpretation	4	3	
6. Assessing the four skills of English			2
7. Selecting assessment tasks			2
8. Implementing assessment tasks			2
9. Evaluating assessment tasks			2
10. Designing assessment rubrics			1
11. Communicating assessment results	1	2	2
12. Conducting research in LTA	1	2	2
13. Assessing native and non-native speakers			1

Using the same method for the analysis of course objectives, a similar pattern of a focus on testing and assessment was identified in course content. CUCs had a dominance of content topics on testing (89.9% = 80/89). CPCs had about two thirds of content topics on testing (61.8% = 34/52) and one-third on assessment (38.2% = 21/52). Similarly, 73.4% of NZ postgraduate course content topics (36/49) were related to assessment, whereas 26.5% (13/49) were about testing. Figure 5 presents the percentage of course content across the three contexts of courses.

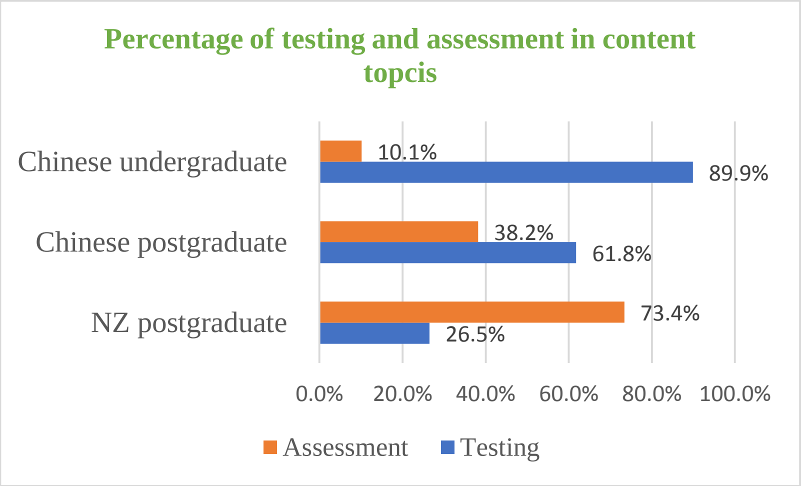

Figure 5. Percentage of testing and assessment in course content topics across syllabi.

The proportion of content topics across the three dimensions of knowledge, principle and skill across all syllabi is presented in Figure 6. The largest proportion of content topics was in the knowledge dimension (46.6% = 88/189), followed by skills (30.7% = 58/189) and principles (22.8% = 43/189).

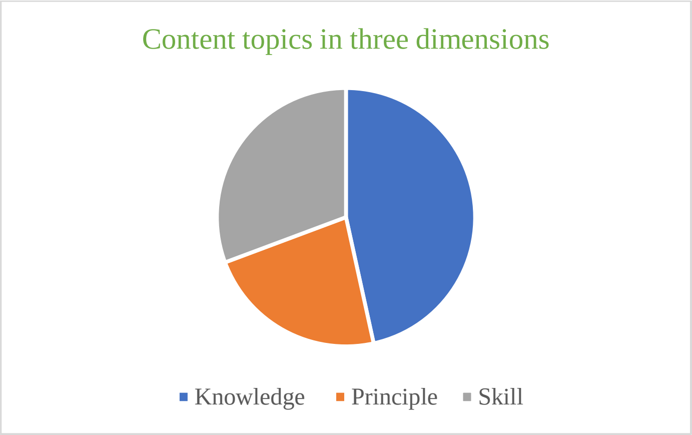

Figure 6. Proportion of content topics in three dimensions.

4.3.1 Chinese undergraduate course content.

Consistent with the focus on testing found in course objectives, the content in CUCs included largely testing-related topics, with a higher frequency of topics in the knowledge dimension than in principle and skill dimensions. In addition, the course content showed a similar structure of how content topics were arranged.

Testing-related topics were the most frequently taught in the ten courses. In the knowledge dimension the following topics were taught with higher frequency than other topics: concepts of testing (9/10 courses); purposes and functions of tests (7/10): types of tests (8/10); testing process (5/10) and test specification (6/10). In the principle dimension the following topics appeared to be important: The relationship between language testing and teaching was taught in six courses; test validity and reliability in five courses; and test validation and test validation in three courses. In the skill dimension, testing different components of English language skills was of importance. Six out of ten courses in this group listed testing listening, speaking, reading and writing as content topics while four listed testing grammar and vocabulary. Item writing and analysis (5/10), tests results interpretation (4/10), and statistical analysis (3/10) also appeared to be important topics.

There was a low frequency of contents related to assessment. Three out of ten courses taught knowledge on formative assessment; two out of ten on portfolio as assessment; one on classroom assessment; one on language assessment principle and one on communicating assessment results. Although half (5/10) course syllabi included assessment related contents (CU1S, CU2S, CU6S, CU9S and CU10S), the assessment topics were restricted to two to four sessions of the entire 8-10 week course period and were usually at the end of the course after an introduction to testing-related topics. The excerpts from the syllabi further illustrated how assessment related contents were taught. UC6S introduced “different approaches to assessment” at the end of the course while UC10S mentioned “classroom assessment” in week 9. UC9S included “formative assessment design” as one of its ten teaching modules after teaching “basic knowledge on tests and its design”. Across the three dimensions of knowledge, principle and skill within CUCs, the largest proportion of topics were in knowledge: 53.4% (47/88) topics were about knowledge; 19.3% principles (17/88) and 27.2% (24/88) skills.

Course content appeared to be similarly structured in CUCs. Courses started usually with teaching knowledge on language testing, followed by content related to principles and ended with content on skills. In teaching knowledge of testing, the topics began with a basic introduction (concepts, purposes, types and history of language testing and test specification) and then moved to the principles, including topics on test validity and reliability, features of a good test, test validation and the relationship between testing and teaching. Content on assessment (e.g., formative assessment, classroom assessment and portfolio) was taught usually after knowledge and principles of testing. Skills of test design (testing listening, speaking, reading and writing, testing grammar and vocabulary, item writing), test evaluation and test/results analysis were introduced usually at the end of the course.

4.3.2 Chinese postgraduate course content.

As reflected in *Figure 6*, though most topics in CPCs were about testing (61.8%), there was a larger percentage of assessment-related topics (38.2%) in CPCs compared with in CUCs (10.1%). Four of the six CPCs incorporated content on assessment (CP1S, CP3S, CP5S, and CP6S). In the knowledge dimension, the concepts of language testing and knowledge on formative assessment knowledge were in a priority: Four of the six courses included these topics. In the principle dimension, similar to CUCs, the relationship between testing and teaching appeared to be an important topic as it was taught by half of the courses (3/6), whereas, unlike CUCs, half of the courses included language assessment principles. In the skill dimension, half of the courses (3/6) focused on topics of testing the four skills of English, statistical analysis and test results interpretation. In contrast to CUCs, teaching content in CPCs focused more on communicating assessment results and conducting research in LTA. In alignment with the postgraduate course objectives, conducting research was taught by half of the courses. Among the three dimensions within CPCs, topics in knowledge were most prominent: The frequency of the content topics about knowledge was 48.1% (25/52); 21.2% on principles (11/ 52) and 30.8% (16/52) about skills.

In course structure, testing-related contents were taught before assessment-related topics and assessment components were taught usually at the end of the courses. The testing-related contents started with a basic introduction to testing knowledge and moved to principles and skills. Assessment contents were taught after the contents of testing which covered between one to five weeks among the entire course length. As evident from the syllabi: CP1S had one week of “activities for formative assessment”; CP5S had four weeks covering “assessment concepts and development”; and CP5S taught “formative assessment in theory and in practice” for one week. CP6S was an exception in that formative assessment was half of the teaching content of the course.

4.3.3 NZ postgraduate course content.

For NZPCs, the content topics were mostly on assessment, more broadly than testing, and showed a greater variety than CUCs and CPCs. Overall, there were 30 types of content topics compared with 24 in CPCs and 22 in CUCs. The topics ranged from various forms of assessment, social dimensions of language assessment, and the process involved in designing and implementing assessment. There was a greater concern with the social-cultural and educational impact of language assessment than in CUCs and CPCs. Topics such as the relationship between LTA and teaching, assessment policy, ethics considerations, assessment and policy and assessing native and non-native speakers were all included. Topics in NZPCs were distributed fairly evenly across the three dimensions of knowledge, principle and skill: 36.7% (18/49) of the overall topics were about knowledge; 30.6% on principles (15/49) and 32.7% (16/49) on skills. However, the organization and structure of how the content topics were taught, differed for each course syllabi.

4.3.4 Summary and comments.

In general, there was an alignment between course content and course objectives. For example, topics of high-frequency in course objectives were also of high frequency in the course content. Similarly, the focus on testing and assessment was similar across the three contexts of courses as found in the course objectives: The CUCs focused mainly on testing-related topics; the CPCs incorporated more content on assessment-related topics than CUCs; NZPGs included the most topics related to assessment. Unlike the emphasis on skills, as identified in course objectives, the greatest percentage of topics across the three dimensions were categorized into the knowledge dimension.

When comparing the two countries, it appeared that, NZPCs viewed assessment more broadly by incorporating multiple assessment methods and putting assessment in social-cultural environment, while Chinese courses (at both undergraduate and postgraduate levels) considered assessment separately from testing and paid less attention to the social dimension of assessment. NZ courses also incorporated a wider and broader range of topics, ranging from the social and educational dimension of language assessment to the relationship between assessment and policy/curriculum. In terms of the content structure, Chinese courses followed a similar arrangement in which language testing was introduced first, before assessment-related topics, with testing covering more of the course period; whereas there was no noticeable pattern of course content structure among NZ courses. In comparing the two academic levels, there was a greater variety of content topics in postgraduate courses than the undergraduate courses: 22 types of content topics in ten CUCs, 24 in six CPCs and 30 in four NZPCs. In addition, research was more emphasized in the postgraduate level.

4.4 Assessment Plans

Assessment plans explain what and how students are assessed during their study. They usually include, but are not limited to, assessment task description, procedures, requirements, marking criteria and contribution to the final grades. Assessment plans support course objectives by providing practical opportunities for students to consolidate the knowledge and the principles taught in the course. Table 9 summarizes course assessment plans with the frequency of each type of assessment task and their average grades weighting (expressed in brackets). Frequency refers to the times certain assessment task appeared while the average grading weight refers to the average of the weighted grade percentage allocated to the assessment task.

Table 9

Assessment Plans by Country and Academic Level

Assessment task		Task frequency and average grading weight		
		CUC (n=10)	CPC (n=6)	NZPC (n=4)
Knowledge and principles				
Final paper-based exam		10 (55%)		
In-class test			1 (20%)	
On-line quiz				1 (10%)
Knowledge, principles and skill				
Presentation		3 (30%)	2 (10%)	1 (10%)
Writing assignment	mid-term homework	7 (30%)		
	mid-term paper		1 (30%)	4 (35%)
	final paper		6 (57%)	4 (55%)
Test design		3 (20%)		
Portfolio of learning			1 (40%)	1 (40%)
Seminar discussion		2 (15%)	3 (13%)	
Others				
Attendance and class performance		9 (10%)	5 (10%)	

The analysis identified nine types of assessment tasks which can be categorized into testing and non-testing tasks in terms of their formality. Testing tasks included final paper-based exam, in-class test and on-line quiz; non-testing tasks included presentation, writing assignment, test design, portfolio of learning, seminar discussion and attendance and class performance. The testing tasks mostly measured knowledge and principles while non-testing assessment tasks were more about the integration of knowledge, principles and skills. Attendance and class performance was also listed as an assessment task. Instead of measuring knowledge, principle and skills of assessment, it was used as a tool to manage, control and regulate students' behaviour.

These assessment tasks were allocated with grade weighting to students' final scores and thus were regarded as summative assessment. In addition to final paper-based paper exam and the end-of-term writing assignment, however, other forms of assessment were conducted during the teaching process which were also regarded as formative assessment.

4.4.1 Chinese undergraduate assessment plans.

There were six types of assessment tasks in CUCs including final paper-based exam, presentation, writing assignment, test design, seminar discussion and attendance and class performance. Test-related tasks was the most important among assessment plans in CUCs, which corroborated the focus on testing as in the course objectives and contents.

A final paper-based exam was used by all the ten courses with the greatest average grading weight (55%). As explained in one of the syllabi statements: "The final exam covered major course content and was usually conducted at the end of the course to check students' understanding. The exam contains questions like multiple choice questions, explanations of terminology and short questions (CU5S)." In addition, test design was an assessment task in three courses with an average weight of 20%. Presentations, mid-term homework and seminar discussions were employed as non-testing assessment tasks during the teaching process. The presentations were in the form of either group or individual work with an average weight of 30%. Students could choose the topics they learned from the course and deliver an oral presentation to the class. Seven of the ten courses assigned mid-term homework as an assessment task (with an average weight of 30%). As described in the syllabi, they were in the form of essay questions and short research reports on topics discussed in the course, mostly related to language testing. All the presentations and the writing assessments were evaluated and marked mainly by the instructors. In one case, the group presentation would be assessed with a combination of "self-evaluation, peer-evaluation and teacher evaluation (CU9S)" rather than the teacher-oriented marking schedule. There was no detailed marking schedule or criteria for presentations and writing assignments in the course syllabi. Two of the ten courses used seminar discussions with an average weighting of 15%.

Students' behaviour was part of the assessment plan in CUCs. Attendance and class performance was recorded and counted in nine out of ten courses as part of their assessment results. As explained in some of the syllabi statements: "The instructor will record attendance and how often students participate in classroom discussions (CU4S); "Participation is important for the course, your attendance and actively answering class questions will be counted as your final grades (CU2S)." Attendance and class performance, however, occupied only an average of 10% of the overall grades.

4.4.2 Chinese postgraduate assessment plans.

There were six types of assessment tasks in CPCs: in-class test, presentation, writing assignment, test design, portfolio of learning, seminar discussion and attendance and class performance. In CPCs, writing assignments were the most common assessment task and student behaviour was also recorded as part of assessment plan.

Unlike the focus of final exams in Chinese undergraduate courses, in CPCs, writing assignments were the most frequently used assessment task (used in all courses) with the heaviest weighting (57%). Generally, there was a detailed description of writing assessments with suggested topics, specific academic purposes, standard formats, recommended referencing procedures and the required word limit (usually around 2000 words). The writing assignment can be divided into three categories: 1) Research reports on interesting issues in language testing and assessment; 2) critical analyses or evaluations of tests or assessment plans; and 3) reports of one's own test development procedures. Some suggested example topics were: "validating a test", "investigating the effectiveness of formative assessment on learning" and "study on washback effect of a language test" (excerpt from CP1 syllabus). Only two of the six course syllabi listed detailed rating criteria for writing assignments (CP1 & CP6).

Two out of six courses listed presentations as an assessment task (with an average weighting of 10%) which were either in the form of individual or group work. The presentation topics could be of interest to students but there was limited information on the format and rating requirements for presentations. One instructor used portfolios (40% of overall assessment) as part of the assessment plan (CP6), which included students' work in progress, for example, self-made tests and writing assignment drafts. Seminar discussion was used by three of the six courses with an average weighting of 13%.

Attendance and performance was also part of the assessment plan in CPCs with an average of 10% contributing to the final grades. As recorded in some of the syllabi: "A rate of excused and unexcused absence will negatively affect the final score (CP5S)" and "students should take an active part in class discussions or answering questions especially when they are required to read the assigned materials before coming to the class (CP2S)".As for CUCs, CPCs used assessment to manage students' behaviour and enhance classroom performance.

4.4.3 New Zealand postgraduate assessment plans.

There were only four types of assessment tasks in NZ postgraduate course syllabi: on-line quizzes, presentations, writing assignments, and portfolios. The major assessment tasks were writing assignments.

Like CPCs, writing assignments were the major assessment tasks listed in NZPCs. All four course syllabi required students to hand in a mid-term paper (accounting for an average of 35% of the overall scores) and a final paper (55%). There were some specific features of the writing assignments in NZPCs. Firstly, NZ postgraduate course syllabi listed very detailed information on the writing assignments including the range of topics, rating criteria, submission methods, referencing requirements, extension and penalties. The topics covered a wider range of topics than the other two groups, corresponding to the greater variety of course content topics in this group. Example topics were: "Review and study alignment between national curriculum and assessment practice in NZ (NZP1S)"; and "a small-scale empirical investigation related to issues of measurement and assessment (NZP2S)". Secondly, more than half of the writing assignments were linked with test/assessment design tasks. Three of the four NZ syllabi required students to design either a test or an assessment procedure. Students were then required to submit a mini-research paper based on their test paper design, or assessment plan development procedure, while referring to relevant language testing theory with appropriate literature support. It was not sufficient to base the review simply on a sample test paper or an assessment procedure. As quoted in NZP1's syllabi,

> The writing assignment for test/assessment development should include an analytical description of the assessment and an evaluation of how good it is for its intended purpose. It should be an assessment that you can obtain adequate information about, and your own involvement in the testing programme and/or an interview with someone else with direct involvement.

4.4.4 Summary and comments.

Assessment plans integrated knowledge, principles and the skills needed for assessment literacy development. The testing tasks measured the knowledge and principles learned in the class while the non-testing tasks measured the integration of knowledge, principles and skills.

In summary, there were different major tasks across CUCs, CPCs and NZPCs. CUCs used a summative final paper-based exam as the major assessment task, which was also given the greatest weighting; Both CPCs and NZPCs used writing assignments as the most frequent assessment tasks with the greatest weighting.

Comparing the two nations, the Chinese courses at both under and post graduate level put attendance and classroom performance as part of assessment plan and linked that with students' final scores, suggesting they used assessment plan as a tool for management and control. However, this was not evident in NZPCs.

Paper-based exams were employed as the most important assessment tasks at undergraduate level, and both Chinese and NZ postgraduate courses used writing assignments as the predominant assessment task. The paper-based exam measured the knowledge and principle dimensions of assessment but could measure practical assessment skills and the high-order thinking skills. It mostly examined students' memory of learned facts and principles rather than the deep understanding of how the knowledge and principles should be applied in practice. Whereas the substantial amount of writing in written assessment tasks provided opportunities for increasing students' thinking and analytic ability as students had to read all the resource materials, synthesize and critically analyse them.

The assessment plans consolidated what was emphasized in course objectives and content and were appropriate for the levels of students. Chinese undergraduate courses focused on the basic skills while postgraduate (both Chinese and New Zealand) required students to demonstrate higher-order thinking/learning skills.

4.5 Chapter Summary and Comments

As Stanny et al. (2014) explained, content analysis of course syllabi can answer a variety of questions about the structure of the course and the focus of teaching. This chapter analysed the general course information, objectives, content and assessment plans of 20 LTAC syllabi collected in two countries (New Zealand and China) across two academic levels (undergraduate and postgraduate).

The analysis revealed an alignment between course objectives, content and assessment plans across the syllabi. Course objectives set the requirements needed to be an assessment literate teacher within the knowledge, principle and skill dimensions. Course content and assessment plans were designed to put course objectives into effect: Course content taught the specific knowledge and principles regarding key issues in language testing and assessment while assessment plans provided students with chances to consolidate the knowledge, principles and to practise the skills required by assessment literate teachers. By the teaching of course content and the implementation of assessment plans, instructors operationalise the course objectives. These course components complemented each other and could be assumed to work together to enhance students' assessment literacy.

Two key findings emerged from the analysis of the syllabi. The first one was the different foci between testing and assessment across different contexts. CUCs had a consistently dominant focus on testing in course objectives, contents and assessment plans; CPCs incorporated other forms of assessment issues in addition to testing while NZPCs viewed assessment in a broader context which included both testing and other forms of assessment. Students were expected to think beyond assessment to understand other related aspects such as the social and educational impact of language assessment.

The second finding was the different focus in the three dimensions of assessment literacy across course objectives, content and assessment plans. In course objectives, skills were emphasized as the most important assessment literacy dimension rather than knowledge and principles. In course content, most content topics were in the knowledge dimension, followed by skills and principles while the assessment plans integrated each of the assessment literacy dimensions of the knowledge, principles and skills.

The emphasis, or lack of emphasis, of course syllabi generally reflects the course instructors' conceptions, beliefs, and their construction of their courses. The initial picture of the LTACs elicited through the analysis of the course syllabi may be similar or different from that of the course instructors. Building on Study 1 findings, the ensuring Chapter 5 probes into LTAC instructors' reflections on how they designed and delivered the courses through a thematic analysis of 20 semi-structured interviews.

Chapter 5 Findings of Study 2: Thematic Analysis of Instructors' Interviews

Study 2 explored LTACs from the course instructors' perspectives as an extension to Study 1. Thematic analysis of instructors' interviews identified themes regarding how instructors designed and delivered their LTACs and how they developed students' assessment literacy. Similarities and differences among instructors were found, leading to a more comprehensive understanding of LTACs construction and implementation and students' assessment literacy development.

The overarching research question in this thesis was: "How are LTACs constructed and implemented in different cultural and academic contexts and how do they shape students' assessment literacy?" To answer this question the following sub-questions guided this study:

1. How do course instructors construct and implement the LTACs?
2. How do instructors seek to cultivate students' assessment literacy?
3. What are the similarities and differences between LTACs in different cultural and academic contexts as reflected in the interviews?

The analysis revealed the similarities and differences between instructors around themes regarding LTACs construction and implementation. To begin with, instructors in both China and NZ used common strategies in teaching, and were confronted with similar challenges. To be specific, instructors in both contexts employed a practice-based approach by taking the context into consideration and teaching through practical tasks. They were confronted with the limits imposed by time and students. There were also differences among instructors: Chinese and NZ instructors held contrasting attitudes to policy and curriculum guidance. Instructors' conceptions of assessment and definitions of assessment literacy also varied, and these individual differences contributed to the different foci in their courses. This chapter reports first on themes of similarity and then move on to themes of differences. Figure 7 displays the thematic map for these analyses.

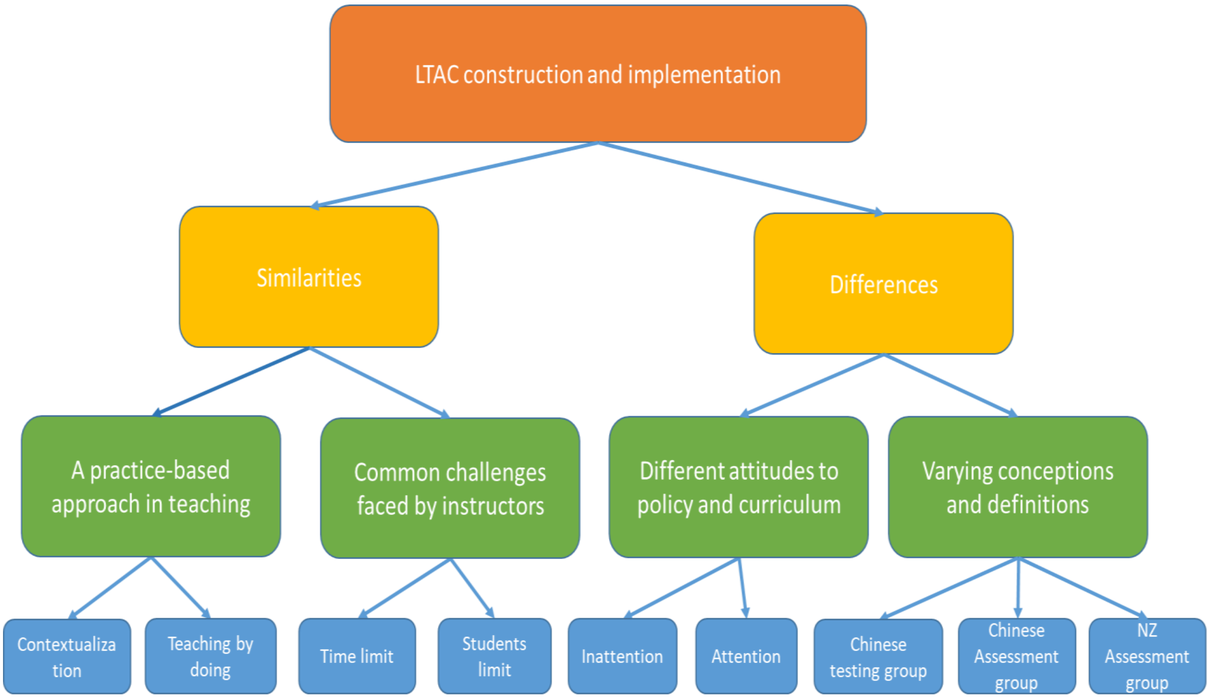
LTAC construction and implementation
Similarities
Differences
A practice-based approach in teaching
Common challenges faced by instructors
Different attitudes to policy and curriculum
Varying conceptions and definitions
Contextualiza tion
Teaching by doing
Time limit
Students limit
Inattention
Attention
Chinese testing group
Chinese Assessment group
NZ Assessment group

Figure 7. Thematic map of instructors' interview analysis.

5.1 A Practice-based Approach

Most instructors (18/20) participating in this study, regardless of whether they were in the Chinese or NZ context, employed a practice-based approach in their course construction and implementation; that is, they prepared students for the teaching profession. Instructors contextualized their courses by building on students' experiences as well as considering their future teaching needs. They also required students to be actively involved in the practical tasks imbedded in the course to help develop assessment literacy.

5.1.1 Contextualizing LTACs.

Although there was an overarching similarity in the practice-based nature of the courses, there were cross-cultural and cross-academic differences because of students participating the courses. In Chinese LTACs, students were learning in a home-country context where they spoke the same language and shared the same culture with the instructor and fellow students. In contrast students in NZ LTACs were from a variety of cultural and language backgrounds and learning in an English language environment.

Students in postgraduate LTACs, in both China and NZ, were either in-service teachers upgrading their professional skills to take up leadership roles or newly graduated from undergraduate education programmes upgrading qualifications to secure better jobs after graduation. Nevertheless, whether they were in-service teachers, or pre-service teachers, students at Chinese postgraduate LTACs in the future would work mostly in China's non-tertiary sectors as ESOL teachers, whereas the students at NZ LTACs would return to their home country (mostly in Asia) or stay in NZ to teach ESOL in various sectors. In contrast, students in Chinese undergraduate LTACs were all pre-service ESOL teachers with no prior teaching experience and little knowledge in language assessment. These pre-service teachers will work in China's non-tertiary sectors such as primary, secondary schools and language training institutes.

5.1.1.1 Taking contexts into consideration.

Considering students' situations, instructors from both China and NZ contextualized their courses to develop assessment literacy needed for their students' future teaching. All the Chinese instructors at both undergraduate and postgraduate levels were aware of the external exam-driven assessment environment and the high stakes of national college entrance exams, thus, they believed they should prepare students for their future careers by teaching them how to design and use tests effectively. This was demonstrated in the following interview excerpts:

> The national exam entrance system decides the importance of tests, a pivotal precursor for students' future life and teachers' career achievement. For a great number of primary and secondary school teachers, they need to teach content required by national English tests and train students with test-taking skills, so tests/testing knowledge and skills were extremely important (CU3 interview).

> Students in my class are pre-service teachers who know little about how to interpret tests and test results, so I would like to teach them how to evaluate test quality and evaluate tests results. That also helped them to design tests, which will help their future teaching (CU4 interview).

> Considering the national examinations teachers need to prepare students for, it's essential for student teachers in my course to understand the sound knowledge of language testing so that they can chose or design the right tests to help their students achieve high scores in the national exam. As a result, I paid a lot attention on the knowledge and skill of language testing (CP5 interview).

Considering the external testing-oriented environment, Chinese instructors tried to teach students to make the best use of tests because they perceived testing as a highly essential and compulsory assessment tool for students' future teaching.

NZ instructors also took students' future teaching into consideration but faced a more complex situation: 1) Students participating the course were from various cultural and educational backgrounds and 2) they would work in a range of teaching environments after graduation. As a result, NZ instructors designed and delivered different versions of the course for specific cohorts and asked students to consider the educational environment in their home country to understand language assessment better.

NZP4 reported that he had accumulated a broad set of articles and resources on language testing and assessment which allowed him to vary course topics and focus according to the contexts students were from, and heading to. He understood that assessment varied according to the contexts, and knew the testing-oriented assessment contexts in Asian countries' culture well. When he had large portion of international students from Asian cultures enrolled in his courses, He would adjust his course to cater to students' home countries' assessment culture by introducing more readings on testing-related topics and asking them how to apply formative assessment in their home cultures. For example, he used to teach a class of Malaysian students. He asked them to read literature about language testing and assessment in Asian countries so they could apply what they learned in the West to their home country. He said in his interview:

> A large majority of international students taking the course are from Asian countries. I used to have a joint teaching programme with other faculty designed for Malaysian students, so I included articles talking about the great impact of exams in learning, something of great interest for students from China, Korea as well. They are pretty well all of the international students and most of them will still work in their home (NZP4 interview).

However, when NZP4 taught students who worked, or would work in NZ context, he asked them to study the NZ schooling system and curriculum as these would be needed for teaching in NZ.

> I do try to be aware if there are NZ teachers working in NZ schools and I will ask them to read more on the NZ schooling system and teaching according to the NZ curriculum requirements. So, I've taken different approaches over the years to the way that I've designed these courses. It varies according to the contexts and depended on the students that I'm dealing with (NZP4 interview).

He claimed that the implementation of formative assessment in the language assessment course should be based on context-specific situations as a general principle for all students. He explained this point in the interview:

> So those are the sort of issues that I discuss and another thing we do, we talked about alternative forms of assessment. What do you think are the principles of classroom-based assessment and standards-based assessment? And I ask them to reflect on the reality of particular contexts and feasibility to implement standards-based assessment or learning-oriented assessment for various conditions including having 60 students in the class or 100 and having in the secondary school maybe 4 or 16 classes of students to deal with (NZP4 interview).

Similarly, NZP2 noticed the increasing number of international students attending her course and made corresponding adjustments. Her course catered not only to local students as previously described, but also incorporated the students' home country assessment environment as well. As she said:

> We are setting up our courses in NZ and targeting mainly for teachers from NZ, however there are more and more international teachers, mainly from Asia joining our courses so we incorporate issues about assessment in exam-oriented environments and relationships between large-scale summative assessment and formative assessment so they know how to use them all (NZP2 interview).

NZP3 also encouraged her students to contextualize their assessment learning within their home country contexts as her students were from all over the world. She gave them freedom to choose a hot topic in assessment, conduct a state-of-the-art literature review and test out the key considerations around the topics. This aroused students' interest and encouraged them to participate actively in these activities. NZP3 remarked in her interview that: "Not surprisingly, students chose topics related with their previous teaching environments with great enthusiasm (NZP3 interview)."

As NZP1 concluded in her interview: "A general principle is that when I designed the course and talked about assessment implementation, I take contextual factors into consideration and I teach students to consider the context as well (NZP1 interview)."

5.1.1.2 The on-line learning system.

In addition, NZ instructors incorporated on-line learning systems in their courses to 1) suit students' learning needs and 2) facilitate students' independent learning off campus. Both NZP1 and NZP4 offered two versions of their courses, an on-campus one and an on-line one but there was not a sharp distinction between the two versions. NZP4, in the interview, said that the on-line course was set up for the convenience of distance students.

> The on-line learning was composed of self-guided reading, practice and forum discussions. Students living in the non-metropolitan area of Auckland can participate in this course for convenience. It's also designed for the expatriate teachers working in China, Korea, and the Middle-east in some cases. While they did their basic study through the on-line study materials, students can have this weekly meeting and distance students who were able to join us at that time with can beam in through ZOOM video conference/communication platform (NZP4 interview).

NZP4 explained that it helped with students' independent learning when they were not able to join the on-site teaching.

> The other thing is that with video we make recording and those who was not able to join us could at least watch the video about the discussion and similarly with the text-chat that used to have a transcript so students who could not join us in the live session could at least read the transcripts (NZP4 interview).

NZP2's and NZP3's courses were taught partially online (in NZP3's case, approximately 50%); students had access to the core and supplementary digital study resources, could contribute to discussion fora and complete online activities and assessment tasks at any time they liked. NZP2 commented: “The on-line learning was helpful for students in terms of theory and factual knowledge learning. Students learned the theory and went through on-line summative tests to check their understanding of the theories and completed required assessment activities (NZP2 interview).”

In summary, although students in NZ and China varied, the instructors took their future teaching contexts and prior assessment experiences into account. In addition, they also provided practical tasks to help develop students’ assessment literacy.

5.1.2 Teaching through practical tasks.

Instructors in both China and NZ believed practice was key to assessment literacy development. They provided as many chances for students to fully comprehend the seemingly complex assessment knowledge and principles and to enhance their assessment literacy.

Most instructors (16 out of 20) emphasized the importance of “practice” throughout the interviews. For example, CU10 said in her interview that: “I teach students the basic concepts, methods and then ask them to design tests. Eventually, it’s the practice they take that help them grasp assessment skills (CU10 interview).” Similarly, CP3 said in his interview, “Although the LTAC I taught is at post-graduate level, I think practice should be integrated with the theory and principles they learned in class. I ask students to learn it by doing. Students also enjoyed this teaching mode (CP3 interview)”; while CP4 stressed the duel importance of theory and practice, “Language testing and assessment is a subject that requires both theory support and practical experiences. So I ask students to practice what they have learned in the class and reflect on theory (CP4 interview).” Likewise, NZP4 said, “The practical application of theory and knowledge is important so we did a lot of practice in the classroom in addition to the essays and paper we wrote (NZP4 interview).”

Building on these beliefs, both Chinese and NZ instructors assigned students test/assessment development projects as important course activities. However, Chinese instructors mostly asked students to design tests while NZ instructors called them “assessment projects”. Eight of the 16 Chinese instructors incorporated test design as major project for their courses and provided relevant activities to help students learn through this process. For example, CP1 asked his students to design and validate a listening test and trialed this test with the freshmen he taught in another class. CU3, CU4, CP2, and CP4 asked students to design a mock test for TEM 4 (China's national Test for English Majors at Band 4), organised group discussions about the mock test and make presentations on this task. CP2 explained this project in the interview:

> I asked students to design TEM 4 reading and writing tests after class applying the knowledge they had learned and then students will discuss their tests in groups under my supervision in the classroom for revision and improvement. Students have to reflect on the strong and weak points in their test design. At the end, they need to deliver a presentation on how they designed the test in front of the class (CP2 interview).

Two instructors, one in China, CP2, and one in NZ, NZP4 were working cooperatively with external organizations in test/assessment development and evaluation projects. They involved students in their own test development projects to help them integrate the knowledge learned in the class and to stimulate their interest in assessment. CP2 invited students to participate in her projects. She said in her interview that:

> I was contracted with some organizations or institutes to design or evaluate some tests, so I will organize students in my class, hold discussion and work together. I would like to involve my students so they can gain the practical assessment skills and that can also greatly arouse their interest (CP2 interview).

NZP4, who was cooperating with a university-based language institute, asked his students to design authentic assessment tasks for the institute. He required students to design, validate, trial and comment on the assessment throughout this process to acquire practical skills necessary to implement the tasks.

Two Chinese postgraduate instructors, CP3 and CP6, asked students to design a classroom-based assessment task to help students develop a more comprehensive understanding of assessment. As CP3 stated in his interview:

> I didn't ask students to design tests for my class, but rather I asked them to develop an assessment task based on the teaching materials I handed out in the class. I believe for postgraduate level, formative and classroom assessment should be put with more emphasis so students can view LTA from a more contemporary and comprehensive perspective (CP3 interview).

Compared with their Chinese counterparts NZ instructors assigned a greater percentage of assessment-related tasks, such as assessment development and evaluation, to help students integrate theory with practice. NZP2 said in the interview that: "In the assessment development task, I encourage students to create formative assessment procedures to help with the learning. Test design is the basic, but I would like them to think beyond the box of testing." NZP4 gave students the freedom to choose between a test design and an assessment development and said he believed that tests or assessment procedures should be context-oriented:

> I want them to have a better understanding to consider what those contexts mean and how to apply them to particular tests, also evaluating the current assessment procedures. If they are in the situation where they have to design their own tests or assessment procedure, they have some ideas and resources to help them do that. They have that sort of tools (NZP4 interview).

NZP2 also asked students to find an ESOL classroom in NZ to observe, and to design an assessment plan for this classroom context. At the end of the course she described how they were required to present their projects to the class.

> They mainly focused on evaluating their assessment tasks that they have done. And by that stage they have already used the strategies with the language learners and they come back evaluating the assessment principles. They need to apply the assessment principles with their project. We look at five principles (validity, reliability, fairness, bias, authenticity) (NZP2 interview).

At post-graduate level, instructors also provided students with the chance to present and display their research or practical work at national or international conferences. CP2 worked together with students on research projects, conferences presentation and publications together to motivate students' learning within the course. She explained in her interview:

> Last year, I spent the entire semester helping one student publish his article in a middle-ranked Chinese academic journal. Even though it's only 300 yuan rewards, it's quite a stimulus for him. In addition, I often select some active students from my class and take them to attend national and international conferences, these are all very good chances for them to learn (CP2 interview).

Through these practice-based approaches to teaching, instructors provided students with chances to practice assessment skills and arouse students' interests in learning, thus developed their assessment literacy. There were also, however, some common challenges in the course construction and implementation.

5.2 Common Challenges

Instructors reported two common challenges confronting them: the time constraints and the students' limitations.

5.2.1 The challenge of limited time.

One of the most common challenges reported by all the instructors was the limited time allocated to LTACs. Chinese instructors had only one semester (from 8 weeks to 16 weeks) to deliver all the teaching and assessment plans on a once-a-week basis, which, according to CU5, was almost "a mission impossible" (CU5 interview). 12 of 16 Chinese instructors wished for more time to deliver the course. CU7 said in the interview that: "One semester is not enough for the teaching. Students need more time to reflect on the knowledge and principles mentioned in the class. The concepts and contents of language testing are not easy (CU7 interview)."

The teaching of formative classroom assessment was even more challenging for Chinese instructors. CP1 stated that: "I wish I could have more time to elaborate on formative assessment, but I have to make a balance because language testing as the current focus. I had only ten weeks (CP1 interview)." CU1 reported that he spent two thirds of the class teaching language testing, and one third on classroom assessment; and that he was trying to overcome the problem of limited time by "incorporating the teaching of classroom assessment in the course and keeping comparing classroom assessment and testing throughout the course (CU1 interview).” The four NZ instructors faced similar challenges of the short time span for the course. For example, NZP1 mentioned that: “One semester is not enough for the course because students need longer period to reflect on the theory and principles of language testing and assessment (NZP1 interview).”

5.2.2 The challenge of students’ limitations.

Most of the instructors (15 out of 20) identified students’ limitations as another challenge. Instructors believed they had to work hard to address: 1) Students’ pre-existing conceptions of assessment; 2) their limited teaching experiences; and 3) other limits such as students’ lack of critical thinking ability. They saw these challenges as obstacles to the development of students’ assessment literacy.

5.2.2.1 Students’ pre-existing conceptions of assessment.

Chinese instructors reported that students were tested with exams throughout their schooling and, therefore, their conceptions of assessment were associated mostly with testing. Their previous repeated testing experiences seemed to reduce their interest in learning about language testing. CU4 claimed this point in his interview:

> Chinese students have been through too many tests that they were not very interested in tests. When I tried to show them how to design and use tests to measure learning, they seemed to be very passive of learning and expressed a message that tests were not really for learning but for entering into college (CU4 interview).

CP5 pointed out that the lack of students’ experience of formative assessment hindered their understanding about the bigger picture of assessment. Learning “new” concepts of formative classroom assessment could be challenging:

> The challenge of the course is from the students. They had been tested at school as the dominating method of assessment and the tests are only for preparing the national exams, so it’s not easy for them to understand the new concept of formative and classroom assessment (CP5 interview).

Likewise CU10 described this notion in her interview:

> Students were preparing for college entrance exams before they entered tertiary study. They were assigned with their teachers' minor weekly tests and major monthly tests to get ready for the final once-for-all national college entrance exam. So their conceptions about assessment is all about tests/testing. It's hard for them to learn other forms of assessment (CU10 interview).

The NZ instructors were also confronted with a daunting task to deal with the diversity of students' culturally imbedded conceptions of assessment. They had a large portion of international students from Asian countries, some students from Latin America and the rest local NZ students. NZ instructors found it challenging helping international students understating "new concepts (as in NZP1 interview)" of assessment. NZP1 pointed out in his interview how cultural background affected students' learning: "Students from Asia are a bit traditional and confined in their own thinking, which makes them difficult to accept the new concepts (NZP1 interview)." Similarly, when teaching students about formative assessment, NZP3 thought students' previous conceptions of testing affected their acceptance and implementation of formative assessment. She commented in her interview that:

> Students' existing beliefs influence their application of what they learned at class. For example I had one student from China who talked about her experience for formative assessment implementation. They want to apply formative assessment but they only experienced testing as a method of assessment so their learning and application of formative assessment can be beyond their thinking. That's the challenges they face and are unable to solve (NZP3 interview).

For NZ instructors, changing students' culturally embedded conceptions of assessment appeared to be challenging but they identified positive prospects. For example, instructors commented that learning in a culturally different nation can be an "eye-opening experience (NZP2 interview)" for international students; the hands-on experience of being a student in NZ educational and assessment system could be a profound influence on their conceptions and practices of assessment. As NZP4 pointed out:

> I want students to have an awareness of what the issues are, there are no simple answers. You've got to balance between current thinking of what represents good

> assessment and practice with realities. You got to establish ways in existing educational system to which assessment is done often by formal tests (NZP4 interview).

5.2.2.2 Students' limited teaching experiences.

Another major challenge came from students' lack of teaching experiences. Most of the instructors (16/20) referred to students' limited classroom teaching experiences as another factor hindering their understanding and application of assessment knowledge and principles.

Instructors at postgraduate level (both Chinese and NZ) reported that students' teaching experiences were insufficient to make a positive impact on their learning of assessment. The teaching of formative classroom-based assessment, which is closely linked with classroom teaching experiences, appeared to be particularly challenging. As CP3's stated in his interview:

> I think one of the challenges in my course is students' limited teaching experiences because the learning of assessment is based on one's teaching and learning experiences. Especially with formative assessment learning, it's more integrated with teaching in the classroom. Students lacking classroom teaching find it hard to understand the concept (CP3 interview).

Even though some students had accumulated several years of teaching experiences, it were inadequate to support their assessment learning. The students in CP1's course were secondary school ESOL teachers with 1-5 years' teaching experiences, but he observed that teachers' past experiences hampered heir learning because they followed the traditional way of teaching. He said in the interview that:

> The teachers in my course have been teaching for some years but their teaching followed their mentor teachers' (practice), the traditional knowledge-transmitting way, which is not that useful for them to understand the concept and method of formative classroom assessment. Assessment learning takes time; it takes a while for students to understand and to apply assessment knowledge and principles into practice, which is best enhanced through teaching practice (CP1 interview).

NZP3, likewise, noted that: “A lot of assessment abilities come through experiences but some students don’t have the experiences necessarily to put the learning into practice. For them that’s another journey (NZP3 interview).” NZP2 also emphasized that the lack of experience held up students’ mastery of assessment principles. She said in her interview:

> The assessment principles are important because they give teachers guidance for why they choose a particular task but they are difficult to get the head around because there is quite a bit of theory behind it which requires a lot of studies and practice with years of experiences. I personally think that they all come down to experiences, the more experienced they are the more they understand why they would choose a particular task and whether it would be suitable for the students. But students lack these experiences (NZP2 interview).

For instructors in the Chinese undergraduate context, the challenge was greater compared with other context as they were teaching pre-service teachers with almost no teaching experience. This is exemplified by CU2’s statement:

> Teaching experiences, or to be particular, experiences in assessing students, hold an important role in understanding these seemingly difficult concepts and terminology in language testing and assessment, but students in these undergraduate education programmes have almost zero teaching experiences, so it’s hard for them to understand this body of knowledge (CU2 interview).

There were other limitations related to the students. Five of 20 instructors commented that students’ lack of critical thinking ability affected their learning. As CU1 noted in the interview: “Students have limited critical thinking ability, especially for those who have never been teaching in the schooling system before. They adhered to old thinking rather than being innovated when facing problems in teaching (CU1 interview).” CP4 also stated that: “Chinese students are usually very quiet, and they dare not challenge their teachers. They spent most of their time listening to teachers rather than producing their own piece of work (CP4 interview).” Likewise, NZP3, suggested that “Students should enhance their critical thinking in order to absorb what has been taught in the class (NZP3 interview)”.

5.3 Different Attitudes to Policy and Curriculum

In addition to the similarities discussed above, the analyses also revealed differences between Chinese instructors and NZ instructors. Chinese instructors at both undergraduate and postgraduate paid little attention to national policy and curriculum guidance on English language education. In contrast, NZ instructors acknowledged the importance of educational policy and curriculum, and incorporated them in their course instruction.

5.3.1 Chinese instructors' response to policy and curriculum.

As described earlier, China's Ministry of Education has released English language education policy and curriculum guidance that advocates using multiple ways of assessment in teaching. However, these policy and curriculum guidance was not considered by Chinese instructors in their course instruction. Of the 16 Chinese instructors, only seven said they had read or learned about the related national policy and curriculum, while13 instructors explicitly expressed that China's national English language education policy and curriculum guidance was not a major concern when designing their courses.

Chinese national policy and curriculum seemed not to be effective and constructive in the eyes of these tertiary instructors, as demonstrated in the following quotations:

> "National policy and curriculum has little impact on teachers' teaching. It's not the concern of the course (CP2 interview)."
>
> "National policy is one thing while teaching in reality is another (CU5 interview)."
>
> "National policy and curriculum doesn't affect my course (CU2, CU3&CU7 interviews)."

Three Chinese instructors relied on their expertise, experiences and students' needs when setting up the course rather than the national policy guidance. For example, CP3's comments about policy in the interview explained: "As a course developer in tertiary education, your course should not be led by national policy. Instead, you should establish the course based on your own learning and understanding of this subject (CP3 interview)." Likewise, even though CP1 participated in the drafting of national English education policy, he didn't take policy and curriculum guidance a factor for consideration in his course design.

He noted in the interview that: “The focus of my course is not based on the policy and curriculum requirement, but more because of my educational experience and my own learning and research work (CP1 interview).” Similarly, CP6 remarked: “We’ve got full control of what happens in our classroom. I design and deliver my course based on what I think students need and what I think they should learn for better teaching (CP6 interview).”

Three instructors thought the policy and curriculum guidance seemed too broad without specific details thus was not effective in their implementation. CU4 explained: “Although the Ministry (Chinese Ministry of Education) advocated formative assessment and classroom assessment methods in addition to testing, it did not provide a clear explanation and concrete examples for teachers to follow (CU4 interview).” CP4 agreed: “National policy is superficial and formalized. It advocates various forms of assessment but fails to provide detailed guidance thus schools still teach and operate in their own ways (CP4 interview).” This was also echoed by CU10: “National policy has not great impact on language teacher education, and the same is true with my course design. The policies are too broad to follow and not applicable to some degree (CU10 interview).”

Three instructors explained it was the college entrance exam that mattered more than the national policy both in the education system and the course instruction. For example, CU4 explained:

> Schools in different levels have their own operating systems but one single goal---sending students to better higher-level educational institutes through the single channel of national entrance exams, so they don’t take national policy into consideration. They take tests into consideration (CU4 interview).

To conclude this section, the Chinese national policy and curriculum guidance was ignored because of the difficulty and ineffectiveness to implement formative assessment in the testing focused educational environment.

5.3.2 NZ instructors' response to policy and curriculum.

In contrast, NZ instructors appeared to place greater importance on policy and national curriculum. They emphasized that students need to know about the national assessment and educational policies and curriculum, which were also part of their course.

For example, NZP3 incorporated language policy as a major course content topic in which she explored the interface between language policies, teaching and assessment, to help students understand the relationship between language policy and language assessment. She argued in her interview that language policy and assessment were highly inter-related:

> We started from the language policy to set the big context to look at the interface between policy and language education so that it has an education focus as well. We look at things like Common European Framework, things like having minimum entry requirement for university or for immigration so students can have the big picture of policy into assessment work. I tried to illustrate how assessment and policy inter-connected, so I look at the social-cognitive framework model, highlighting to students that all of the different aspects of assessment that really impact each other and the inter-relationships there (NZP3 interview).

NZP2 reported that she was aware of NZ's national language policy and curriculum. Having studied it on the Ministry of Education assessment website, she shared it with her students. She was aware, however, of the uniqueness of NZ policy and its function, stating in her interview that:

> The national policy and curriculum are catering to the NZ context, so I told my students that they need to think about the corresponding national policy when designing assessment activity either for this course assignment or for their future teaching (NZP2 interview).

NZP4 asked his students to read documents about language policy as background learning materials, whereas NZP1 asked his students to review the NZ national curriculum and policy as assessment guidance. NZP1 said he thought highly of NZ national policy in terms of assessment:

> I think NZ educational assessment policy is the best amongst the world because of the small number of teacher and students in NZ, it allows the educational policy to go more in depth and detail. NCEA (National Certificate of Educational Achievement) is composed of half internal assessments and half external assessments, which allows students greater chances reaching their academic achievement goals. Instead of teaching to prepare for exams, NZ schools tend to teach knowledge that students think is interesting and helpful for their future (NZP1 interview).

5.4 Varying Conceptions and Definitions

In addition to the contrasting attitudes to policy between instructors from the two nations, the analysis also identified differences among instructors regarding their conceptions of assessment and definitions of assessment literacy, which had direct impact on their course design and delivery. Based on these differences, the instructors in Study 2 were categorized into three groups: Chinese testing, Chinese assessment and NZ assessment group. The Chinese testing group believed testing was the major form of assessment and thus their courses centred solely on language testing. This group included seven instructors: five undergraduate (CU 3, 4, 5, 7, 8) and two postgraduate (CP2 and CP4). The Chinese assessment group believed assessment should be more than testing and their courses incorporated components of formative and classroom assessment, formative assessment and classroom assessment. This group included nine Chinese instructors: five undergraduate (CU1, 2, 6, 9, 10) and four postgraduate (CP1, CP3, CP5, and CP6). All four NZ instructors had similar conceptions of assessment and used assessment as an umbrella term that included both testing and other forms of assessment and addressed a greater variety of topics in their courses. Thus, they were called the NZ assessment group.

5.4.1 Conceptions of assessment.

In the interviews, the 16 Chinese instructors used the Chinese word "*ceshi*" (testing) to refer to exams or tests and the word "*pinggu*" (assessment) for non-testing assessment that often used in formative classroom-based assessment. They viewed "*ceshi*" and "*pinggu*" separately as two independent methods of assessing students. "*Pinggu*" was regarded as a newer concept; a more recent method which functioned as an alternative to testing. For the

Chinese testing group, their conceptual framework of assessment was represented by one circle of testing. The Chinese assessment group's conceptions of assessment had two circles of testing and assessment, assessment being new, emerging, but smaller compared with testing. In contrast, NZ instructors viewed assessment as a big circle that incorporated testing and other assessment methods. Figure 8 illustrates the conceptual frameworks of assessment of the three groups of instructors. The following sections reports the conceptions of assessment and definitions of literacy in each group respectively.

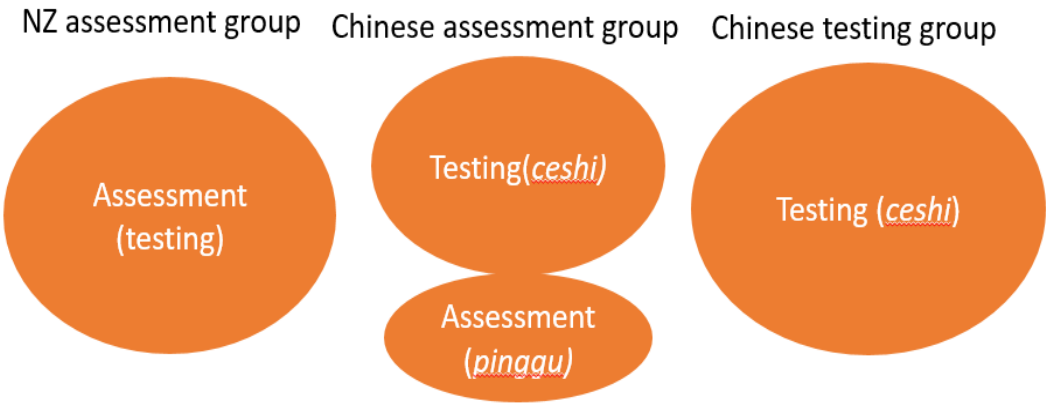

Figure 8. Instructors' conceptual framework of assessment.

5.4.1.1 The Chinese testing group.

The Chinese testing group regarded testing as the major and the only way of assessing students' learning. Their conceptions of assessment appeared to be equivalent to conceptions of testing. They believed in, and supported the dominant role of testing as the only assessment method to teach and assess students in China; they argued for the importance of tests. These views were evident in the following interview excerpts:

> "Tests remain and will remain the major form in English language teaching in China (CU4 interview)."
>
> "Knowledge of language testing should be taught to students because that's how teachers know about students' learning and that's a skill or technique greatly needed by teachers (CU7 interview)."
>
> "I only focus on testing in my course for I believe this is one of the most important assessment tools for teachers and the students (CP2 interview)."

5.4.1.1.1 Testing as the tradition.

Four of seven instructors in the Chinese testing group mentioned it has been a long-lasting tradition for LTACs in China to put the sole focus on testing, thus they followed the traditional way of teaching: “This course is relatively traditional by focusing on the various types of language tests with explanation on concepts like test validity and reliability (CU4 interview)”; “We’ve always set the content of the course as language testing (CU5 interview)”; “Teaching testing is the tradition. It’s like a necessity for pre-service teachers to learn how to test and use test results to help students to learn (CU7 interview)”; “I was assigned with the task of teaching “language testing” which was set up by previous teachers under institutional instruction, so I continued with the teaching of testing (CU8 interview).” These instructors played safe by following the tradition of teaching testing because the course had been set up in this way either by previous instructors or by the faculty.

5.4.1.1.2 Assessment should be in pedagogy.

Four instructors in the Chinese testing group claimed that *pinggu* (assessment) should be more closely linked with teaching and should be taught in a pedagogy course, or learned through students’ future teaching. Specifically, they believed that formative assessment and classroom-based assessment were more about teaching rather than assessing. For example, CU5 noted in the interview that:

> Formative assessment is associated with teaching. It’s an on-going but spontaneous system which could be an observation of a student or a question and answer. It’s not the concern during the course for it is to be learned in pedagogy courses (CU5 interview).

Likewise, CP4 believed formative assessment should be part of the teaching done by instructors in other faculty. She said the interview that: “I think assessment should be taught in language teaching pedagogy courses. These courses are usually set in the faculty of education (CP4 interview).” Their belief in the similarity between assessment and pedagogy made it hard for instructors in the Chinese testing group to teach assessment in their LTACs.

Two instructors believed formative classroom assessment should be learned in students' future teaching rather than in the course because they had too limited teaching experiences to understand assessment. CP2 mentioned the near-zero experience of students as a barrier for implementing formative assessment, saying that: "Most of the students in the course had little teaching experience, so it's difficult to teach them formative assessment, which I think should be learned and enhanced from their teaching practice (CP2 interview)."

5.4.1.2 The Chinese assessment group.

Unlike the Chinese testing group, the Chinese assessment group incorporated non-testing assessments into their courses and viewed assessment from a broader perspective, but they viewed testing and assessment as two separate concepts. As CU9 explained in his interview: "I think testing and assessment are different in that testing is the paper-based exam that most students will have to do at the end of the semester, while assessment is a newer and broader concept (CU9 interview)." CP5 agreed testing and assessment were different with testing as being stagnant while assessment being dynamic.

> I think assessment is different from tests that usually happen at the end of the term. It's a dynamic process that required teachers to know students well and the role of teacher and students were exchangeable. Whereas, tests are usually pre-set and stagnant. Teachers can adjust assessment constructs and even ask students to make assessment criteria with them, which make the classroom assessment more subjective compared with testing (CP5 interview).

Nevertheless, no matter how they defined assessment, they all regarded it as an important assessment method. They believed assessment could be used to counteract the negative washback effects of testing, to improve teaching/learning and, if applied well, could enhance students' dispositional growth.

5.4.1.2.1 Assessment counteracts the negative washback effect of testing.

Six out of nine instructors in the Chinese assessment group explicitly expressed their concern about the negative washback effects of the traditional testing culture in China. For example, CP1 said in the interview that: "Teaching-to-the-test is a long-lasting prominent problem in our educational system, be it in primary, secondary or tertiary sectors. It is highly inefficient and poses harmful effects to learning and teaching (CP1 interview)”. CU6 expressed a similar opinion: “A lot of teachers in China have realized the negative washback effect of testing, especially how that affects one’s cognitive, dispositional and emotional development (CU6 interview).”

Unlike the Chinese testing group’s adherence to the tradition of testing, the Chinese assessment group broke with convention by considering formative classroom assessment as a solution to the testing-only environment. They believed this could help bring about change by offsetting the negative testing effects. CU6 would like to break the constraints of testing and he explained in the interview that formative classroom assessment can help measure students’ communicative English ability that cannot be measured by the current tests.

> Testing can exert negative effects on students’ learning because of its limits in assessing students’ overall ability. The final tests scores cannot represent students’ English communicative ability, so I incorporated assessment in my course with the hope that my students could learn to use formative assessment to help students improve practical English ability and to reduce the harmful effects of testing (CU6 interview).

CU10 claimed that the testing culture has caused social problems, thus assessment should be a way to address these problems. He said in his interview that:

> Testing used to but is no longer the one and only important factor in assessing language learners for it causes a lot of social problems. Schools, parents and students put too much emphasis on the final tests scores, which twisted our educational system. Teachers in China are not cultivating students to be able learners, but rather test machines, so they need to understand that examination is not the only assessment method for students’ learning (CU10 interview).

CP5 believed assessment could help improve learning and be reflected in students’ test results.

> Assessment itself should be promoting learning so a teacher should understand what assessment is, how and when to use non-testing assessment to apply that in

> classroom teaching. In this way, you are helping students to learn English rather than pass the tests. I believe if formative assessment were implemented in the right way, it should be helpful in improving students' final grades while improve learning (CP5 interview).

These instructors were aware of the negative washback of testing and tried to reform their courses by incorporating assessment so that student teachers would learn testing was not the only way of assessing students. They felt that formative classroom assessment would be a powerful and useful alternative.

5.4.1.2.2 Assessment to improve teaching quality.

Three instructors argued that, because teachers spent much of their time teaching, formative classroom assessment should be used more frequently as a way to ensure the quality of teaching. As CU10 explained in the interview: "Assessment was more useful for school teachers. They are assessing in the classroom more often than with the paper tests (CU10 interview)." CU1 and CP3 held similar conceptions that classroom assessment has greater effect on teachers' teaching:

> School teachers have little rights to exert when dealing with the tests in their teaching. They should follow either the school or the head teachers in terms of the tests, but they enjoy greater freedom as they choose which classroom assessment tools to apply. Plus, they spend quite a lot time in the classroom, so I think formative assessment is more important for teachers (CU1 interview).

> Teachers should work out how to assess students in the classroom with non-testing methods for they have little control with the final entrance exam. I would like to teach them how to apply observations, feedback and self-assessment in the teaching so that they can take good use of their classroom teaching time, which I believe is more useful than teaching students how to do the tests (CP3 interview).

These instructors claimed that as teachers had little control over the designing and administration of tests, learning test development may not be of great help for their future teaching. Instructors believed student teachers should learn how to design and implement formative assessment activities in classroom to help improve teaching and students' learning.

One instructor, CP1 taught students to use tests in a diagnostic and formative way. CP1emphasized classroom assessment by teaching students how to use and analyze tests in a scientific and effective way to improve students' learning. He explained to them that "Testing without evaluation is ineffective (CP1 interview)." In his words, he "teach (es) the concepts of testing when thinking about teachers' teaching in the classroom. (CP1 interview)". He asked students to think about the formative application and function of a test and to think about the following questions: "What's the objectives of the test? How did the test help students improve their learning? How to use the information of the tests in the teaching? And what are the washback effect of testing?"

5.4.1.2.3 Assessment to facilitate dispositional growth.

In addition to helping with teaching and learning, two instructors viewed assessment as a useful method to facilitate students' dispositional growth. CU2 talked about the negative effect of testing on students' mental health and argued that formative assessment should be a tool to help students grow mentally stronger. She explained in the interview that: "Testing has exerted great anxiety on students, parents and teachers. Teaching-to-the-tests neglected students' emotional and social needs. That's why I added formative assessment in my course to provide a chance for students to grow as a person (CU2 interview)."

CU1 thought testing diminished students' interest in learning and turned them into passive learners. He claimed in his interview that formative classroom assessment could help students grow as motivated and capable learners by saying that:

> Given China's current schooling system where test results are the greatest concern, teachers should prepare students for the exam so to present a good score to either please the principal or the parents. But that's more like an external motivation, I would like my student teachers to find the inner drive in assessment. I want them to know that formative assessment can improve students' ability and inner drive in learning. I think it's very important for their teaching. As a teacher, the autonomy and the ability to learn in your students should be your greatest achievement (CU1 interivew).

CP6 echoed this opinion in her interview asserting that: “Formative assessment requires students’ efforts, so they can learn to be responsible in measuring their own learning progress and thus manage their learning (CP6 interview).”

5.4.1.2.4 Influences on instructors to incorporate assessment.

The four instructors in this group (UC1, UC2, CP3 and CP6) attributed the breakdown of the testing-only tradition in their courses to their learning about formative assessment. CU1 acquired his doctoral degree with a research focus on language testing but, after he started his teaching career, formative assessment was introduced to his institute as part of an educational reform. His senior colleagues organized an assessment learning group and invited him to join the learning community, thus ushering him to the world of formative assessment. He started to gain familiarity with, and confidence in using, formative assessment and, consequently, added it to his own language testing and assessment course. His conception of assessment changed after his work with the formative assessment learning group.

> I unconsciously equated testing to assessment in the past as I believed it was the only way of evaluating students’ English ability. But after the learning of formative assessment, I realized their differences and realized that assessment was more important for my student teachers. Now I realized that formative assessment was more useful for primary and secondary school teachers because they are assessing in the classroom more often than paper tests (CU1 interview).

Similarly, language testing had been a long-established traditional course in CU2’s university but, based on her teaching reflection and her study experience in the U.K., she added elements of formative assessment in her LTAC. In her interview she said that: “My learning experience about formative assessment in the U.K. was eye-opening. I started to incorporate formative assessment and other ways of assessing students in my teaching. I would like to teach students this new concept (CU2 interview).”

CP3 is the director of a language assessment reform and research centre in China. He regularly attended international conference on language assessment and was aware of the latest trends in this field. With what he had learned from international conferences, he reflected on his teaching and made corresponding changes. His said in the interview:

> I've been learning language testing for more than a decade but in the international community of language testing and assessment, researchers and educators have long realized that language testing is limited. Assessment is not only about testing and testing is not the only way of assessing students' language learning. I participate international conferences every year. In this year's LTRC (Language Testing Research Colloquium, 2016), the conference theme was "From language testing to language assessment, linking teaching learning and assessment". That indicates the trend from language testing to assessment. As I read from the latest literature, I also learned the benefits of other forms of assessment and I feel obliged to teach teachers to use other forms of assessment to improve students' learning (CP3 interview).

Likewise, CP6 has been teaching LTACs for more than a decade, which has helped her to reflect and develop her knowledge on assessment. Furthermore, her interest and study in classroom assessment after learning "the Assessment Reform Group" (Black & William, 1999) project inspired her to upgrade her course from traditionally testing-only to include formative classroom-assessment. She said in her interview that:

> Previously my course was called language testing and was mainly about testing. Three or four years ago, I grew more interested in language assessment. I read Professor Dylan William's "the classroom experiment" and was fascinated by the reform. After that, I learned more about classroom assessment and I started to divide my course into two parts: language testing and classroom assessment. I believe classroom-based assessment can promote better learning if used in a right way. It should help students to improve learning and test scores at the same time (CP6 interview).

To conclude, though community of learning, through individual reading of cutting edge literature, through individual overseas learning experiences and international conference participation, instructors from Chinese assessment group learned a broader concept of assessment and gradually changed their conceptions of assessment, which helped them to reform their LTACs.

5.4.1.3 The NZ assessment group.

NZ instructors viewed language assessment as a big circle that incorporated testing and other forms of assessment. The word "assessment" was used throughout in New Zealand instructors' interviews, functioning as an umbrella term to indicate testing or other forms of assessment. NZ instructors did not distinguish testing and assessment as separate concepts. In NZP4's interviews, language testing was a form of assessment. When he assigned students with the task to develop an assessment, it could either be a performance-based assessment or a written reading test.

> It's possible to choose and look at a speaking assessment that is not necessarily a test, but of course depending on what you define as a test…There are formative elements like the university course here that are two of three components contribute to the grade for the course, which is a summative test. But generally, I think assessment is broader and testing is part of assessment (NZP4 interview).

NZP2 believed testing was more formal and controlled while formative classroom-based assessment could be informal but still equally needed: "A test is something more formal with controlled conditions. You get students to produce something. You complete certain tasks and you analyse it, whereas informal assessments can be done on the run and many schools do it in a classroom (NZP4 interview)." She gave equal weighting to different forms of assessment while commenting on New Zealand teachers' assessment practice. "The New Zealand teachers are doing informal formative assessment all the time. But I think there should be a place for both. So, in this course, I teach both testing and other forms of assessment (NZP4 interview)."

NZP3 argued that the distinction between formative and summative assessment can be artificial and that a test was not, necessarily, a summative assessment if it provided information on the students' progress and informed teaching and learning.

> I think it can be an artificial categorization or distinction between summative and formative assessment. There can be more formative design of tests and formative assessment can operate in a summative way. So, I teach students very much about

> good assessment. There's no point in assessing unless it has a formative dimension, I guess that's one of the messages that I tried to give my students. The results need to be given to students as feedback in teaching and learning and into practice (NZP3 interview).

The NZ assessment group valued formative feedback based on clear criteria when students were completing their assessment tasks. NZP2 emphasized the importance of feedback during teaching by saying in the interview: "Feedback is important. Students send us sections of their projects (mostly in writing) and we give them feedback on that. They get formatively assessed in the programme themselves in this sense (NZP2 interview)." Likewise, NZP3 valued discussion in the classroom with her students, or among students themselves, as part of learning to assess through completing assigned assessment tasks. She helped students prepared for their assignment through class discussion and thinking that built towards their final writing. As she said in the interview:

> The assessment focus is very much linked with what we are doing in class and in the module. So in class it will be a combination of me talking, presenting information, a lot of peer work, a lot of group work, a lot of discussion sometimes a debate sometimes activities like students might do a solo presentation on an article that they read… and all these build towards their final assignment on assessment with the hot topics they chose. The classroom is quite vibrant. No final exams. They have three assignments to do, one for each module. All the assignments are kind of what we tease down and thought about but obviously they need students to think about in their spare time as well (NZP3 interview).

5.4.2 Definition of assessment literacy.

The definition of assessment literacy also varied across the three groups of instructors. Chinese testing group thought assessment literacy was equal to testing literacy. Chinese assessment group held similar definitions with NZ assessment group that assessment literacy incorporated more.

5.4.2.1 Chinese testing group.

The definition of an assessment literate teacher held by the Chinese testing group instructors was associated largely with testing knowledge and skills. When they were asked about what made an assessment literate teacher, they described a teacher with the ability to design, administer and evaluate tests. Assessment literacy in their definition is similar to testing literacy.

As CU3 stated in his interview: “In order to be assessment literate, student teachers should learn the basic concepts, theoretical and statistical knowledge about language testing.” More specifically, it is to have “the ability to choose the right test and appropriately interpret tests results (CU4 interview)”; it is to have “the ability to design a valid test and choose the right tests for students (CU7 interview)”; it is to have “the ability to come up with a test that measures what is taught in the class (CU8 interview)”; and it is to possess the ability to “integrate testing theory with practice to design and evaluate English tests (CP4 interview)”. As summarised in CU5’s interview:

> Assessment literate teachers are those who have learned the four components of language ability listening speaking reading and writing. In this way, they can design the appropriate tests for each component. And then he or she should be clear about the different types of the questions. Once an assessment literate teacher gets the tests, he or she should possess the ability to measure what the test examines. In addition, he or she should be able to measure the quality of the tests from the perspectives of validity and reliability (CU5 interview).

5.4.2.2 Chinese assessment group.

The Chinese assessment group, when asked about their definition of an “assessment literate” teacher, first acknowledged the importance of “testing literacy”, and then pointed out the value of other forms of assessment. They believed an assessment literate teacher should be aware of multiple ways of conducting assessment with students. As CU1 said in his interview that:

> An assessment literate teacher should master the basic knowledge and skill of testing, say how to design and evaluate the tests. In addition to testing he or she should also know how to use other forms of assessment like classroom assessment and formative assessment. The former is necessary in China's educational environment while the latter can help to increase students' internal motivation to learn (CU1 interview).

They viewed testing and assessment differently, and thought assessment literacy should be testing and assessment literacy. The Chinese assessment group also set higher standards for assessment literacy compared with the Chinese testing group claiming there were other qualities which are essential elements of assessment literacy. Three instructors emphasized knowing the purpose and context of assessment. This "why to assess" should be the first step of any quality assessment for assessment literate teachers. As CP1 asserted in his interview: "An assessment literate teacher has to first of all hold a clear purpose or objectives, so during our course, we remind our teachers of their teaching purpose."

As well as knowing "why to assess", members of the Chinese assessment group also emphasized knowing "what", "why" and "how" to assess as part of assessment literacy. CU9 pointed out the importance of knowing "what" in his interview: "An assessment literate teacher should know the content, or to be specific, the language point or ability to be assessed (CU9 interview)". CP3 echoed this point in his vision of an assessment literate teacher from the interview: "My conception of an assessment literate teacher is the one who is able to work out exactly what they are trying to assess based on their students' ability and course objectives (CP3 interview)." CP5 pointed the importance of understanding what components to assess: "As a language teacher, one should learn about language learning progression for specific groups of students and put assessment requirement into a reasonable range accordingly (CP5 interview)."

Then, they also pointed out that assessment literacy included the skills needed to assess students, the "how to assess". CP3 argued for this ability in the interview that: "Assessment literate teachers should possess the skills of assessment. They should have the knowledge and being able to carry out appropriate assessment activities to assess what they've learned in the classroom (CP3 interview)."

5.4.2.3 NZ assessment group.

When asked about what an assessment literate teacher should be, the NZ assessment group had a similar definition with the Chinese assessment group: An assessment literate teacher should be able to understand the "why", "what" and "how" to assess. For example, NZP1 emphasized the "why and what to assess" in his interview that:

> An assessment literate teacher should know what this assessment is for, whether this assessment is designed by him or herself or chosen from other resources. Before he set out using this piece of assessment, he should be aware of the purpose of implementing this work and what exactly is measured (NZP1 interview).

NZP2 agreed, pointing out the reason for assessing: "They should know the reasons why they are assessing. They've got to think of the reasons why they are doing the assessment and then think about what actually they are going to assess (NZP2 interview)."

NZP4 emphasized the importance of "how to assess" and have appropriate knowledge and practical skills needed by an assessment literate teachers and they set that as the course outcomes:

> They need to know the current practices so to carry out effective assessment. For example when we talk about speaking assessment, we look at design options like classic IELTS type interview, as compared with a peered format, we look at the Cambridge exam or the small group discussion but also we talked about direct assessment versus and semi-direct, computer-based assessment as you get in the TOFEL exam (NZP4 interview).

The NZ assessment group also identified the importance of ethical considerations in assessment, which had not been noted by either of the Chinese groups. NZP1 included ethical considerations in "how to assess" by arguing that an assessment capable teacher should report assessment ethically as well as effectively.

> Another important aspect is about ethics. When assessment results come out, how can teachers better communicate assessment outcome to students without being demoralizing them while providing the correct feedback? They should think about that questions and interpret the assessment results ethically to students (NZP1 interview).

5.5 Chapter Summary

The findings of Study 2 extended that of Study 1 and added to the understanding of LTACs from the instructors' perspectives. The similarities found included: 1) instructors from all contexts contextualized the courses according to students' backgrounds and the external contexts; and 2) instructors designed practical tasks to help develop students' assessment literacy during the courses. This practice-based teaching reflected the focus on skills in course syllabi objectives as identified in Study 1. The instructors in all contexts believed that these strategies would enable students to become assessment literate for their future teaching in specific contexts. Furthermore, Chinese instructors were affected largely by the external testing culture while NZ instructors took students' cultural backgrounds into consideration. In addition, all instructors were confronted with the challenges of time constraints and students' limitations.

One of the identified differences concerned the instructors' attitudes to policy and curriculum guidance. Compared with New Zealand instructors, Chinese instructors' (at both undergraduate and postgraduate levels) attitudes towards national policy and curriculum appeared to be reluctant, passive and unconcerned. They constructed their courses based on their knowledge, experiences and students' needs. New Zealand instructors, on the other hand, recognized the importance of policy and curriculum guidance by incorporating them as part of the courses and asking students to learn to assess according to policy and curriculum.

Another difference found between instructors concerned how they conceptualised assessment and how they defined assessment literacy, which further affected their course construction and implementation. Based on these differences, instructors were categorized into: Chinese testing group, Chinese assessment group and NZ postgraduate group. The Chinese testing group regarded testing as the major focus for assessment, while the Chinese assessment group acknowledged the importance of formative and classroom assessment. Both groups accepted the current testing-oriented environment but applied different strategies. The testing group focused on teaching the knowledge and skills of language testing so that students could design and evaluate tests for better reliability and validity. In contrast, the assessment group advocated the importance of formative and classroom assessment and taught students how to apply it in classroom teaching to counter the negative washback effect of tests and testing. The NZ assessment group regarded assessment as a bigger circle that incorporated language testing and other forms of assessment.

This categorization of instructors was different from that of Study 1. In Study 1, instructors were categorized into Chinese undergraduate, Chinese postgraduate and NZ postgraduate according to the cultural and academic contexts and different course foci were found across the three contexts. Chinese undergraduate course had a dominant focus on testing; Chinese postgraduate courses had more elements in assessment compared with Chinese undergraduate while NZ postgraduate courses had assessment as an umbrella term incorporating testing. The differences between this categorization indicated instructors' individual differences.

Based on the findings of Study 1 and Study 2, there emerged a tension between testing and assessment, especially within the Chinese undergraduate context. In addition, the external assessment environment exerted great influence on the construction and implementation of the course and thus may cause impact on students' assessment literacy development. In order to explore the effectiveness of LTACs on students' assessment literacy development, Study 3 looked into the Chinese undergraduate context and examined the assessment literacy development of student teachers, namely, the senior Chinese undergraduate pre-service teachers (PSTs) in a web-based survey (section 3.4.2 Phase-II quantitative data collection.also explains the reason for recruiting PSTs as survey participants). The following Chapter 6 reports the findings of Study 3.

Chapter 6 Findings of Study 3: Statistical Analysis of Students' Survey

Informed by Study 1 and Study 2 findings, Study 3 surveyed Chinese undergraduate senior pre-service ESOL teachers (hereafter PSTs) with the intention to: 1) illustrate a conceptual model of PSTs' assessment literacy by looking at three components: conception, self-efficacy and self-reported practice of assessments; 2) identify the structural relationships among the three components and 3) see how LTAC participation relates to students' assessment literacy by comparing LTAC participants with non-participants. After data cleaning and normality check, confirmatory factor analysis (CFA) measuring each component was first conducted on the theoretically-informed structures and three measurement models were established. Based on these models, structural equation modelling (SEM) was conducted to identify the relationship between the three measurement models and also to identify the inter-relationship between latent factors. Then multi-group invariance testing was performed according to LTAC participation condition and factor mean scores were compared in ANOVA for factor invariance.

In accordance with recommendation on SEM report by Schreiber et al. (2006), the report of the quantitative analysis addressed the six non-technical issues: 1) stating and explaining research questions; 2) a rationale for the use of statistical analytical tool (explained in section 3.3.3 Questionnaire for students survey., Chapter 3); 3) the theoretical framework of the model (explained in section 2.3.3.2 Measurements on conceptions and practices of assessment., Chapter 2); 4) listing sufficient information on descriptive statistics in the form of tables, figures and texts; 5) providing a graphic display of the final models (the path coefficients in CFA and SEM models were all displayed in standardized regression weights); and 6) providing discussions and implications concerning the findings.

The overarching research question in this thesis asked: "How LTACs are constructed and implemented in different cultural and academic contexts and how they shape students' assessment literacy?” This study addressed the second half of the question, namely, “how LTACs shape students’ assessment literacy”, thus the following two sub-questions were then designed:

1. What are PSTs’ conceptual model of assessment literacy components and how do they inter-relate?
2. Does taking an LTAC have an impact on PSTs’ assessment literacy?

6.1 Participants and Data Cleaning

The web-based online survey collected a total of 1533 responses (from 25 tertiary institutes across eight provinces in China) out of 3068 visits at a response rate of 50.06%, higher than the average on-line survey response rate of 31.8% (Crawford, 2002). Data cleaning was performed before descriptive analysis for valid demographic information.

Since all the items were set as forced response, no missing value was reported, thus the following steps were applied to ensure data quality for further analysis.

Step 1: Participants who skipped all the items were removed (namely, 173 (11.93%) out of 1533 responses).

Step 2: Entries with a response time faster than 207 seconds (at least 3 seconds for one item with 69 questions in total) were deemed invalid rapid responding (Sue & Ritter, 2012). As a result, 291 (21.39%) out of 1360 responses were deleted with 1069 remaining.

Step 3: Participants who gave the same response on each item from the beginning to the end were searched and deleted, thus a further nine participants were deleted with 1060 remaining.

To conclude, a total 473 responses out of 1533 (30.85%) were deleted which resulted in the final 1060 valid responses. Given that this questionnaire included 63 items as manifest variables, the response to variable ratio met the requirement for desired case-to-variables ratio (5:1) analysis (Field, 2016). Participants’ demographic information is presented in Table 10.

Table 10

Participants' Demographic Information

Demographic	n	%
Gender		
Female	962	90.78%
Male	98	9.22%
Year		
Year 3	585	55.24%
Year 4	475	44.76%
LTA or no LTA		
Yes	296	27.92%
No	764	72.08%
Types of LTAC		
Language testing	99	9.34%
Language testing and assessment	151	12.25%
Others	46	4.34%

6.2 Instrument

The instrument used for this survey was the Conception, Self-efficacy and Practice of Assessment (CSEPoA) questionnaire that measures three components of assessment literacy: conceptions, self-efficacy and self-reported practices of assessment. CSEPoA is composed of a demographic section including six questions and three sub-instruments including 63 statements in total. The three sub-instruments are: Conceptions of Assessment in Chinese Context (C-CToA), which contains seven latent factors with 30 items (Brown et al., 2011), Self-efficiency of Assessment (SEoA), which includes one factor with eight items (Gu, 2016) and Practices of Assessment (PoA) which examines three factors with 25 items (see Appendix D for the full version of this questionnaire respectively in English and Chinese). Information about the development, translation, piloting and validation of CSEPoA can be found in section 3.3.3 Questionnaire for students survey., Chapter 3. Table 11 lists the instrument latent factors with their definitions.

Table 11

CSEPoA Latent Factors and Definitions

Factor	Definition
Conceptions of Assessment in Chinese Context (C-CToA)	
Student Development (SD)	Assessment is for student development or betterment. Assessment cultivates positive moral and ethical qualities, and values in students which contribute to their lifelong and life-wide learning, and good citizenship. A wide variety of valued personal and social skills appropriate to full participation in society are developed. The goal is to help students develop positive social conduct, moral character, and appropriate personal potential and qualities.
Help Learning (HL)	Assessment is to help students learning. Assessment is a means of improving the quality of both students' learning and teachers' instruction. A variety of assessment techniques are used to identify the content and processes of student learning, as well as the quality of instruction.
Accuracy (AC)	Accuracy indicates that assessments are accurate and reliable. Assessment correctly measures students' ability and performance with accurate report or scores. The assessment results are trustworthy and dependable.
Irrelevance (IR)	Irrelevance means assessment serves no legitimate role within teaching and learning. While assessments may be administratively required, teachers' knowledge of students based on long relationship and their understanding of curriculum and pedagogy precludes the need for assessment. Since accurate and precisely correct measurement of assessment is difficult, teachers may have legitimate grounds to ignore assessment.
Exam (EX)	Examination (EX) emphasizes the effect of examination as a major assessment tool in China. Assessment holds students accountable for learning what was expected of them by society. This is usually done using performance on examinations or tests. This requires grading, scoring, or evaluating student performance against standards, objectives, targets, or expectations.
Error (ER)	Teachers should consider measurement error when using assessments and reporting assessment results. Assessment results may not represent students' authentic ability or performance because of measurement errors.
Teacher and School Control (TSC)	Assessment is used managerially to control schools or classrooms. The assessments are not necessarily scored or recorded, rather they lead to better discipline. Assessment is used to enhance and maintain the control of the teacher and the dominance of teachers' opinions over those of the student.
Self-efficacy of assessment (SEoA)	
Self-efficacy (SE)	Self-efficacy refers to the belief in one's innate ability to carry out a task in a proper and appropriate way. Self-efficacy of assessment means feeling certain about the assessment knowledge and skills one

	has in assessment (including both testing and non-testing assessment methods), also the ability to analyse, interpret and communicate assessment results to parties involved.
Practice of assessment (PoA)	
Testing Practice (TP)	Testing practice refers to the assessment tasks and activities that related to test, which is usually conducted periodically at certain point or at the end of teaching. Examples of testing practice include test design, administration, evaluation and test results interpretation and communication. Its meaning overlaps with summative assessment but the emphasis is on using testing as a form of assessment.
Formative Practice (FP)	Formative assessment practice refers to the assessment tasks and activities that teachers conduct during teaching in the classroom. When incorporated with classroom instruction, it provides information on how to improve teaching and learning so that timely adjustments can be made. Formative practice in this thesis emphasized on the non-testing assessment methods such as checking students' understanding, learning portfolio, self-assessment, peer-assessment with emphasis on feedback and learning goals.
Exam Preparation (EP)	Exam preparation refers to the assessment tasks and activities that help students prepare for exams, especially the high school and college entrance exams (*zhongkao/gaokao*). The purpose of the assessment is all about good results and exam techniques. Exam practice emphasized using assessment for exam preparation regardless of its form and these practices include teaching exam skills, teaching to the test with focus on what tested in the exam, assess against exam rules and using formative assessment for exam purposes.

Note. The definitions are adapted and adjusted from Brown et al., (2011).

6.3 CFA Model Analysis

In this section, CFA model specification procedures and results are reported. Univariate and multivariate normality with all items were checked. Then CFA was performed separately with the three sub-instruments using Maximum Likelihood (ML) calculation (Fabrigar et al., 1999) based on the theoretically-informed structure.

6.3.1 Normality assumption check.

Item level univariate normality was checked with skewness and kurtosis. The cut-off range of skewness and kurtosis in this study were respectively at Kline's (2015) standards of skewness <|3.00| and kurtosis <|8.00|. The results were within the cut-off range, showing that each item is normally distributed. Detailed results for univariate normality calculation results of each item can be found on Appendix E.

Multivariate outliers were checked with Mahalanobis distance and probabilities calculation with *df* (Ullman, 2006). With this calculation, 127 outlying participants who had probability values below 0.001 were identified. However, Osborne and Overbay (2004) suggested within large samples, outliers can be kept for further analysis. A solution is to make comparison between models with and without outliers to determine whether removing outliers can make better model fit or provide more valuable information. The Akaike information criterion (AIC) was used to identify superior fit when lower AIC values indicated better fit (Foldnes, Foss, & Olsson, 2012).

6.3.2 Measurement model for C-TCoA.

The measurement model testing for C-TCoA was based on two models established in previous research: Model A, a first-order model with seven factors (as seen in Table 11) and Model B (as shown in *Figure 9*), a second-order model with three meta factors covering the seven first-order factors (Brown et al., 2011). The three meta-factors included Improvement, Accountability and Irrelevance. Improvement consisted of Student Development, Help Learning, and Accuracy; Accountability contained Examinations, Error, and Teacher and

School Control; and the last meta-factor was Irrelevance. Figure 9 shows the simplified schematic graph of this second-order model.

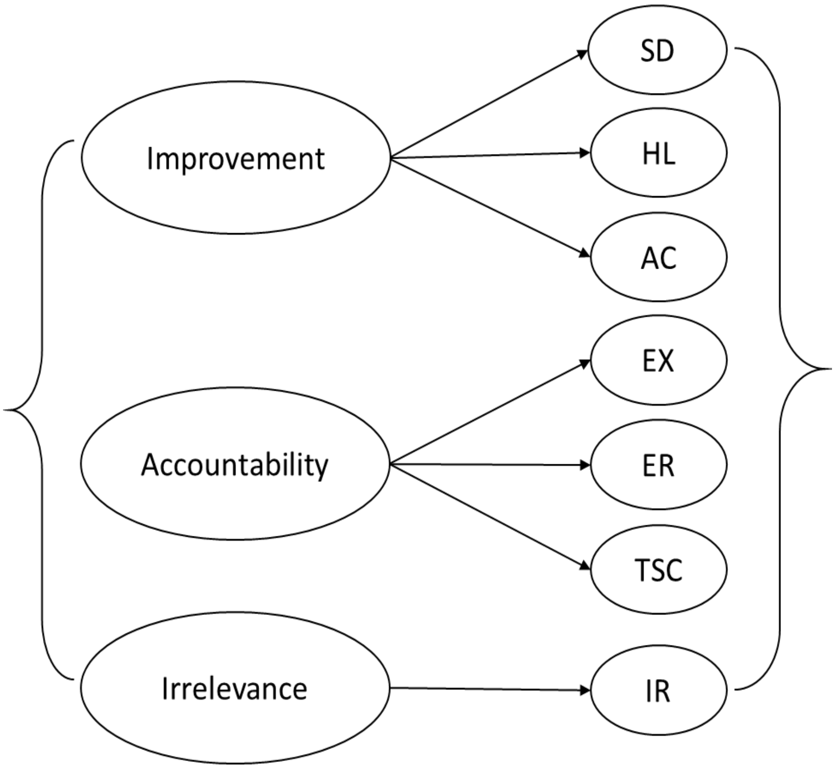

Figure 9. Simplified schematic model for second-order C-TCoA CFA model.

Note. SD = Student Development; AC = Accuracy; ER = Errors; IR = Irrelevance; EX = Exam; HL = Helping Learning; TSC = Teacher School Control. The curlicue brackets indicate that the factors are correlated; the items were removed for simplicity.

CFA was first run separately to evaluate model fit of Model A and Model B both with and without outliers, but neither of the models showed the presence of acceptable model fit (model fit criteria are listed in Table 5, section 3.5.3.3, Chapter 3). Although for all the models the RMSEA, SRMR, and gamma hat values met the criteria of acceptable fit, the $\chi 2/df$ values exceeded the acceptable level of 3.8 while the CFI values were below 0.90. This indicated the PSTs surveyed in this study did not hold similar conceptions with previous surveyed samples (mostly in-service primary school teachers). Thus all the models (Model A, A1, B and B1, see Appendix G) were rejected for the partial acceptance and better fit model was yet to be tested.

Compared with Model B, Model A, the first-order seven-factor model showed model fit, thus model modification was carried out on this seven-factor structure. model modification principles mentioned in section *Note*. CFI=comparative fit index;

RMSEA=root mean square error of approximation; SRMR= standardized root mean residual; AIC=Akaike information criterion.

3.5.3.4 Model modification.Chapter 3, items with low cross loadings and large modification indices were removed one by one to test model fit with and without outliers until a fitting model was established. In total, 5 items (EX7, EX4, EX5, HL2, TSC4) were deleted before acceptable model fit was found. Models with no outliers showed better fit and lower AIC values thus the final trimmed CFA model with no outliers (Model G1) was kept for further analysis. Appendix F lists the item code, item statement, and the values used to select the item for deletion. Appendix G lists the model fit indices for each step of model modification. Figure 10 displays the schematic graph of the final trimmed CFA model for C-TCoA.

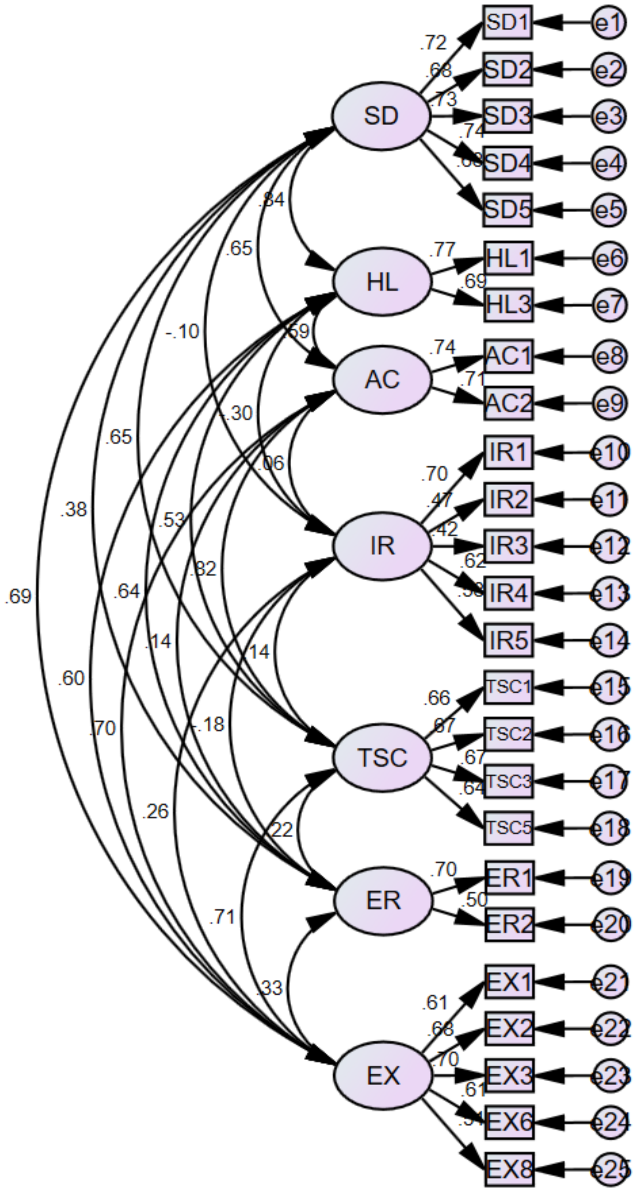

Figure 10. Graphic figure for the trimmed CFA model of C-TCoA (Model G1).

Cronbach's alpha values for each factor were calculated in SPSS for internal reliability validation while inter-correlation values were calculated in AMOS for the relationship between facotrs. The average factor means, and standardized deviation values were also calculated in SPSS for analysis. Negatively worded items in this sub-instrument (IR3 and IR4) were reversed prior to factor mean score calculation. The results are reported in Table 12.

Cronbach's alpha values for each factor were all above 0.70, indicating acceptable internal reliability (DeVellis, 2003). The highest mean scores among the seven factors were that of Error, which suggests PSTs agreed the most that measurement errors should be taken into consideration and assessment should be used cautiously, whereas the lowest mean scores were that of Irrelevance, suggesting PSTs agreed least that assessment should be irrelevant and ignored.

The highest correlation among the factors was between Student Development and Helping Learning (*r*=0.84), indicating that when PSTs believed assessment improved students' development, they also agreed that assessment helped students' learning. Accuracy and Teacher School Control also had a strong positive relationship (*r*=0.82) suggesting that PSTs associated accurate assessment as a tool to control schools and teachers.

Exam was an important factor that had three strong positive relationships with other factors, namely, Student Development, Accuracy and Teacher School Control. This indicated that PSTs strongly agreed that exam 1) should help students' development, 2) should be accurate and 3) should be used by teachers and schools to control teaching and learning. Three negative relationships existed between Irrelevance and Student development; Irrelevance and Helping learning; Irrelevance and Error. This indicated that PSTs disagreed assessment was irrelevant.

Table 12

Descriptive Information for C-TCoA Factors

	Inter-correlations (*r*)								
C-TCoA model	1	2	3	4	5	6	7	M	SD

factors									
1.Student Development	(.83)							3.50	0.80
2.Helping Learning	.84	(.70)						4.06	0.85
3.Accuracy	.65	.59	(.70)					3.01	0.77
4.Irrelevance	-.10	-.30	.06	(.71)				2.46	0.60
5.Teacher School Control	.65	.53	.82	.15	(.75)			3.15	0.77
6.Error	.38	.64	.14	-.17	.22	(.73)		4.22	0.86
7.Exam	.70	.60	.70	.26	.71	.33	(.76)	3.12	0.73

Note. Values larger than 0.70 in bold; factor estimate of reliability, Cronbach's alpha in brackets.

6.3.3 Measurement model for SEoA.

CFA analysis of the one-factor, eight-item sub-instrument of SEoA produced acceptable model fit except for very high value for the chi-square to *df* ratio both with and without outliers (Model SA and Model SA1, see Appendix G). Inspection of the modification indices suggested removal of 1 item, SE1 (see Appendix F for this item details). The trimmed seven-item single factor model showed consistently good fit both with and without outliers (Model SA-SE1 and Model SA1-SE1), but Model SA1-SE1 was kept for further analysis due to better model fit and lower AIC value (see Appendix G for details). Figure 11 presents the schematic figure of SEoA.

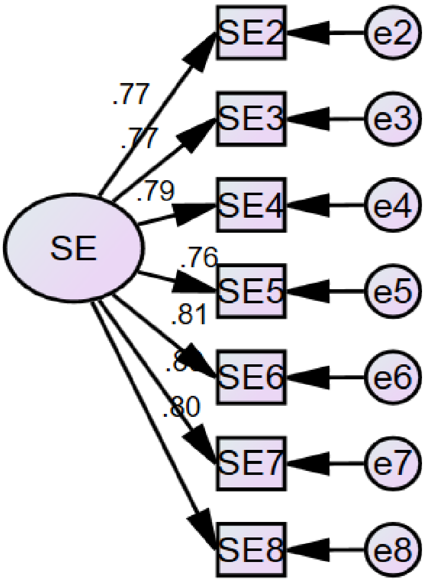

Figure 11. Graphic figure for the trimmed CFA model of SEoA (Model SA1-SE1).

The factor estimate of internal reliabilities was good (Cronbach's Alpha=0.92) and the factor mean score for Self-efficacy was 3.52 (SD =0.71).

6.4.4 Measurement model for PoA.

The PoA sub-instrument was developed based on Study 1 and Study 2 findings (see section 3.3.3.1 The development of CSEPoA.) and CFA was run on the three-factor structure (formative practice, testing practice, and examination preparation). The original models were rejected because of poor fit both with and without outliers (Model a and Model a1). Then model modification was carried out with the best intention to keep the three-factor structure. Inspection of factor loadings identified two items with low standard regression weights (TP4 and TP6). Thus, these two items were deleted first to test model fit. However, the model fit indices were below the acceptable range, thus modification indices were inspected and unfit items were deleted one by one for good model fit. After deleting another 7 items (TP2, TP5, EP2, EP3, FP5, FP9, FP10, see Appendix F for item details), the trimmed model with no outliers (Model j1) was kept for further analysis with acceptable to good fit and lower AIC. Appendix F lists the item code, item statement, and the values used to select the item for deletion. Appendix G shows the model fit for the model modification process and Figure 12 shows the schematic graph of PoA CFA model with standardized loadings.

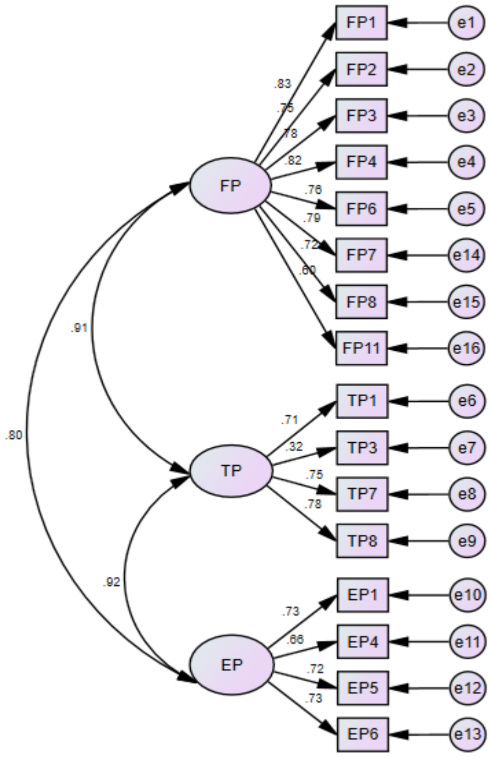

Figure 12. Graphic figure for the trimmed CFA model of PoA (Model j1).

The latent factor means, Cronbach's alpha values and inter-correlation values were also calculated for PoA (see Table 13). Cronbach's alpha values of all factors were above 0.70 which indicated acceptable internal reliability. Among the three factors within PoA, the highest factor mean scores were that of Formative Practices, indicating PSTs' strongest agreement of assessment as formative whereas the lowest factor mean scores were that of Testing Practice, indicating their weakest agreement with testing related tasks and activities. The three latent factors were all highly positively correlated. This suggested that PSTs conceived these factors as strongly similar and strongly linked with each other.

Table 13

Descriptive Information for PoA Factors

	Inter-correlations (*r*)				
PoA model factors	1	2	3	M	SD
1. Formative Practice	(.91)			4.38	0.73
2. Testing Practice	0.91	(.73)		4.00	0.67
3. Exam Preparation	0.80	0.92	(.80)	3.99	0.73

Note. Values larger than 0.70 in bold; factor estimates of reliability Cronbach's alpha displayed in brackets.

6.4 SEM Model Analysis

After the three CFA measurement models were established, the conjoint SEM structural models were tested with all participants (with no outliers) based on the following three possible assumptions: 1) Conceptions and self-efficacy were correlated and they work together to predict practices of assessment (Model 1); 2) Self-efficacy functioned as a mediator between conceptions and practices (Model 2) and 3) Conceptions predict practices and then predict self-efficacy (Model 3). The best SEM model was chosen for better model fit and lower AIC value. AIC calculates the sum of negative log-likelihood and a penalty term that increases with the number of parameters in a given model. The negative log-likelihood indicates the goodness of fit for a proposed model while the penalty terms indicates the model complexity. The minimal AIC value indicates the best balance between model fit and model complexity and thus model with the smallest AIC is regarded as the best model among competing models (Lin, Huang and Weng, 2017).After the best SEM model was established, multi-group SEM was conducted for invariance testing between LTAC participants and non-participants.

6.4.1 Establishing SEM model with full sample.

Based on the three possible structures, Models 1, 2 and 3 were tested with SEM and Table 14 shows the model fit results which were all within good range, though the differences between each index were small. To decide the best models using AIC, the following four steps were conducted: 1) the differences of AIC between each model were first calculated, 2) exponent of (-0.5*difference) were calculated 3) all weights were summed and 4) the proportion of sum for each model were determined to choose the best model.Using that we can see Model 2 had 99% of the AIC weight and was smaller than Model 1 by 9.88 points and Model 3 by 104 points, providing convincing evidence of support for Model 2 (Burnham &

Anderson, 2004), thus Model 2 was kept as the best SEM model. In this model, self-efficacy functions as a mediator between conceptions and practices of assessment.

Table 14

SEM Model Fit Results for Best Model

					Model fit indices				
Model	χ2	χ2/df (*p*)	CFI	RMSEA	SRMR	Gamma hat	AIC	ΔAIC	AIC weight
Model 2	2238.898	2.184 (0.000)	0.942	0.036	0.0437	0.95	2536.898	--	.99
Model 1	2238.898	2.184 (0.000)	0.942	0.035	0.0437	0.95	2546.780	9.88	.01
Model 3	2258.780	2.189 (0.001)	0.942	0.035	0.0431	0.91	2640.898	104.00	.00

Note. CFI=comparative fit index; RMSEA=root mean square error of approximation; SRMR= standardized root mean residual; AIC=Akaike information criterion.

Predictor paths were calculated from each factor and statically non-significant paths were removed. After removing statistically non-significant paths between factors, the new model still had consistently good model fit (χ2=2261.969; χ2/*df* =2.169; CFI=0.942; RMSEA=0.035; SRMR=0.0443; gamma hat=0.96). Figure 13 displays the simplified graphic representation of the SEM model with significant paths only.

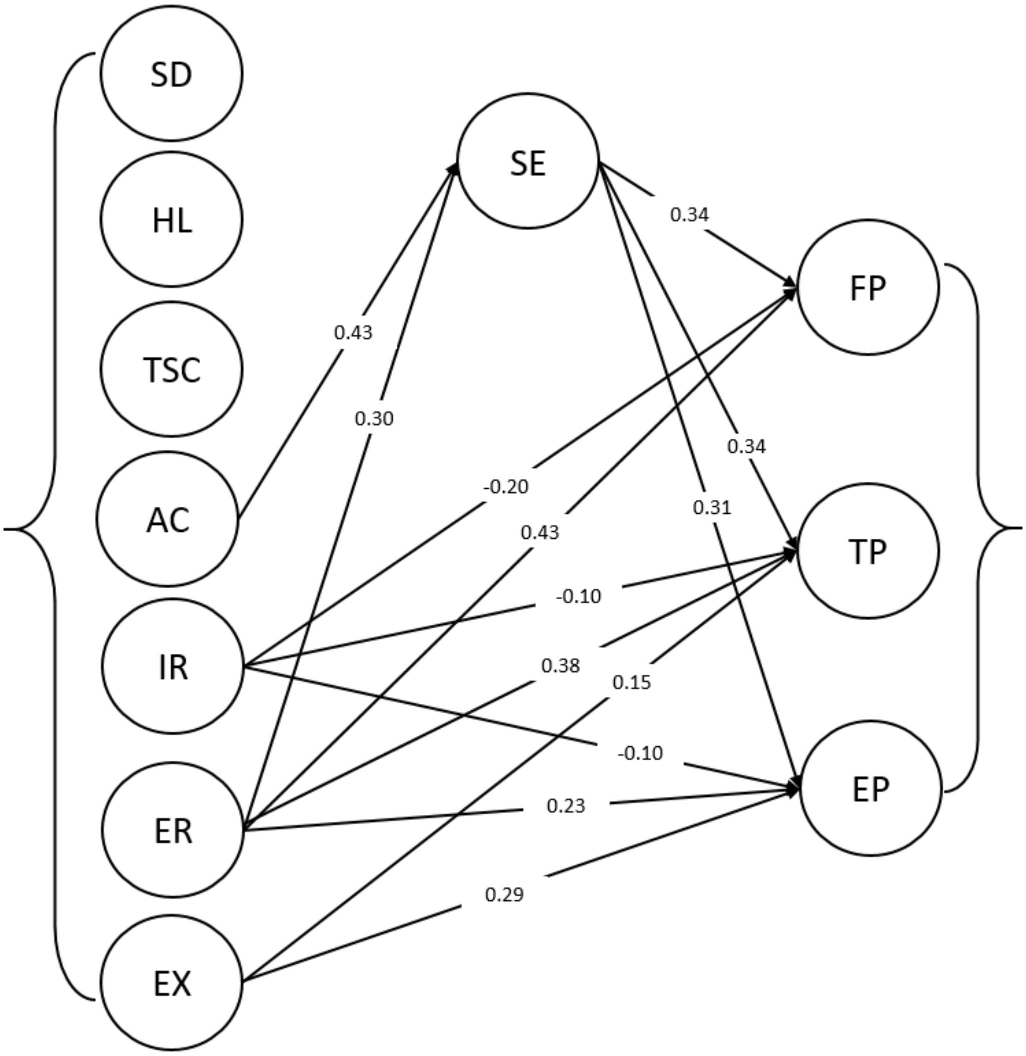

Figure 13. Simplified schematic graph of SEM model to full sample.

Note: SD = Student Development; AC = Accuracy; ER = Errors; IR = Irrelevance; EX = Exam; HL = Helping Learning; TSC = Teacher School Control; SE = Self-Efficacy; TP = Testing Practice; FP = Formative Practice; EP = Exam Preparation. The curlicue brackets indicate that the factors are correlated; the items were removed for simplicity; the values are standardized.

In total, there were 13 significant paths between the 11 latent factors: Two factors in conceptions of assessment (Accuracy and Error) positively predicted assessment Self-efficacy but all to the moderate degree. This indicated that when PSTs regarded assessment as accurate, and when they took good consideration of assessment errors, they had good confidence to carry out assessment activities.

The three paths from Self-efficacy to the three factors of practice of assessment were all significant with similar values, though the influences were not strong (all below 0.40), suggesting self-efficacy of assessment almost equally predicted the three forms of assessment practices.

There were eight significant paths between three conceptions of assessment (Irrelevance, Error, and Exam) and the three assessment practices. Irrelevance weakly and negatively predicted all three forms of assessment practices (regressions all below 0.20). This indicated that when PSTs thought assessment as irrelevant, they preferred not to use the three assessment practices. Error positively predicted all three practices, suggesting when PSTs agreed that that awareness of error in assessment influenced all three assessment practices, especially the formative assessment. Exam weakly and positively predicted Testing Practice and Exam Preparation but didn’t predict Formative Practice. This indicated when PSTs believed assessment was for exams, they carried out testing practice and exam-preparation related practices but no formative practices, which seems consistent with Chinese norms.

6.4.2 Invariance testing.

In this section, invariance testing as nested multi-group SEM was conducted for LTAC-participants and non-participants to test whether LTACs made a difference to the established SEM Model 2 structure. Following the traditional sequence of equivalence testing in multi-group SEM (Vandenberg & Lance, 2000), the analysis established:

1. Configural equivalence: the factor structure keeps the same across groups;
2. Metric equivalence: factor loadings are equivalent across groups;
3. Scalar equivalence: all intercepts of item loadings on factors are equivalent;
4. All regressions from factors to other factors are equivalent;
5. All covariance between inter-correlated factors are equivalent;

The equivalent residuals are not required for determining invariance. In accordance with the current conventions for assessing measurement invariance, the decision rule of ΔCFI ≤ 0.01 was used to define invariance because the chi-square test is overly sensitive to large sample sizes (Cheung & Rensvold, 2002; Wu, Li, & Zumbo, 2007). Based on these criteria, full SEM model was strictly equivalent between LTAC participants and non-participants (see

Table 15 for model fit results). This indicated that whether PSTs in this survey participated in LTACs or not made no statistically significant difference in their conceptions of assessment, their self-efficacy for assessment, or their practices of assessment.

Table 15

SEM Model Fit Results for Equivalence Tests

	Model fit indices				
Model	$\chi2/df$	RMSEA	SRMR	CFI	ΔCFI
1. Unconstrained (configural)	1.676	0.027	0.0584	0.934	
2. Measurement weights (metric)	1.663	0.022	0.0590	0.934	0.000
3. Measurement intercepts (scalar)	1.667	0.027	0.0596	0.935	0.001
4. Structural weights	1.665	0.027	0.0620	0.932	0.003
5. Structural covariance	1.56	0.025	0.0620	0.932	0.000

Note. CFI=comparative fit index; RMSEA=root mean square error of approximation; SRMR= standardized root mean residual; AIC=Akaike information criterion.

6.5 Factor Mean Comparison in ANOVA

Since the multi-group SEM established full invariance in the model, factor mean scores were calculated and compared using ANOVA together with Cohen's effect size (*d*) to examine the difference LTAC participation made on each factor. The factor mean scores were calculated using the popular regression method (Odum, 2011). According to Hattie (2008), in education research, the effect size is trivial when the absolute values of $d \leq 0.20$; small when $0.21 \leq d \leq 0.39$; moderate when $0.40 \leq d \leq 0.59$ and large when $d \geq 0.60$.

The factor mean score comparison results are listed in Table 16. As can be seen from the table, there were significant differences in factors mean scores for Helping Learning, Error, and Self-efficacy, Formative Practice, Testing Practice and Exam Preparation between LTAC participants and non-participants. This suggested that LTAC participants agreed more than can be explained by random processes on the following ideas compared with non-participants: 1) assessment helped learning and 2) measurement errors should be taken into consideration when doing assessment; LTAC participants also had higher level of self-efficacy of assessment and higher level of acceptance of all three forms of assessment practices. However, the Cohen's *d* values for these factor mean comparisons indicated the effects were either trivial or small, suggesting that insofar as mean scores are concerned, having an LTAC made little difference.

Table 16

Latent Factor Mean Score Comparison

	LTA or Not	Mean	SD	F	*p*-value	Cohen's *d*
Assessment conceptions						
Student Development	Yes (269)	0.419	1.090	0.660	0.417	0.424
	No (674)	-0.017	0.962			
Helping Learning	Yes (269)	0.188	1.059	13.534	**0.000**	**0.259**
	No (674)	-0.075	0.966			
Accuracy	Yes (269)	-0.005	1.068	0.009	0.924	0.007
	No (674)	0.002	0.973			
Irrelevance	Yes (269)	-0.091	1.013	3.121	0.078	0.127
	No (674)	0.036	0.993			
Teacher School Control	Yes (269)	0.057	1.057	1.224	0.269	0.079
	No (674)	-0.023	0.976			
Error	Yes (269)	0.213	0.988	17.327	**0.000**	**0.301**
	No (674)	-0.085	0.993			
Exam	Yes (269)	0.022	1.055	0.179	0.673	0.030
	No (674)	-0.009	0.978			
Self-Efficacy						
Self-efficacy	Yes (269)	0.134	1.051	6.800	**0.009**	**0.241**
	No (674)	-0.009	0.978			
Assessment Practices						
Formative Practice	Yes (269)	0.242	0.979	22.491	**0000**	**0.343**
	No (674)	-0.096	0.993			
Testing Practice	Yes (269)	0.180	1.015	12.321	**0.002**	**0.252**

	No (674)	-0.072	0.986			
Exam Preparation	Yes (269)	0.119	1.046	5.329	**0.021**	0.164
	No (674)	-0.047	0.978			

Note: p-value <0.05 in bold, Cohen's *d* > 0.21 in bold.

6.6 Chapter Summary

This chapter reports the analysis results of Study 3, the students' survey. It first established three measurement models for sub-instruments of C-TCoA, SEoA and PoA and then explored the inter-correlation between latent factors within each measurement model. It is found that among the factors in T-CoA, Student Development and Helping Learning had strong positive relationships, and so did Accuracy and Teacher School Control. Student Development and Helping Learning are more related to the purposes of formative assessment and that explained why they were positively linked. Accuracy and Error are concerned with the validity and reliability of assessment, the results indicated that PSTs cared about the accuracy of assessment results, which may be due to the high stakes of the entrance exams. Exam is related with summative testing and it had positively strong relationships with Student Development and Teacher School Control. This finding corresponded the Chinese school tradition that exam was used to cultivate students' personal growth and to manage and control school and teachers (Chen & Brown, 2013).

After the establishment of measurement models, a conjoint SEM model was established and confirmed one of the assumption that self-efficacy functioned as a mediator between conceptions and practices of assessment. Insignificant paths between factors were removed for further analysis. The results indicated that students cared much about assessment accuracy and errors and believed it was a major factor that determined their assessment practices. The results also showed PSTs are more confident to conduct assessment when they make sure assessment is accurate and when they have taken errors into consideration.

The multi-group SEM revealed that full invariance was established for the effect of LTACs. Latent factor mean scores comparison showed there were six statistically significant differences with LTAC participants in Helping Learning, Errors, Self-efficacy, Formative Practice, Testing Practice and Exam Preparation.

Study 3 probed into the Chinese undergraduate context with a focus on three assessment literacy components: self-efficacy, conceptions and practices of assessment of pre-service ESOL teachers. The findings revealed self-efficacy served as a mediator between conceptions and practices of assessment. LTAC participation enhanced pre-service teachers' self-efficacy of assessment though it had little impact on the overall assessment litearcy component model. These findings shed light on teacher education in other examination-oriented societies like India (Brown et al., 2015) or the Arabic world (Gebril & Brown, 2014). The mediating function of self-efficacy between conceptions and practices reinforced the well-established importance of self-efficacy and thus could be a focus area in teacher education programs. If student teachers' self-efficacy for assessment were improved, their conceptions and practices should be strengthened correspondingly. In addition, though there was little difference between LTAC participants and non-participants in terms of their assessment literacy component model, course participation has been shown to be effective in enhancing self-efficacy of student teachers (Dunlap, 2005; Palmer, 2006). This also proved the importance of assessment course for teacher education programs and student teachers' assessment literacy development.

Chapter 7 Discussion

This chapter begins with a conceptual summary and discussion of the findings as a coherent whole. Then it elaborates on the theoretical and practical implications informed by the findings. Limitations of the thesis and suggestions for future research are presented afterwards, and this chapter concludes with the major contribution of this thesis.

7.1 Review of Major Findings

This review of major findings is organised around four aspects:

1) A model of influences generated from findings of the three studies.
2) The contextual factors that affected course syllabi, instructors' conceptions and students' assessment literacy development.
3) The impact of instructors' individual differences on their courses and
4) The limited effect of the assessment course on students' assessment literacy.

7.1.1 The model of influences.

Threading the numerous findings together, a pattern of how various contextual factors shaped LTACs and helped develop student teachers' assessment literacy has been identified. Rather than report the findings in a list-wise fashion, a model of assessment literacy development in assessment preparation courses as being applicable in various contexts is generated and reported (see Figure 14).

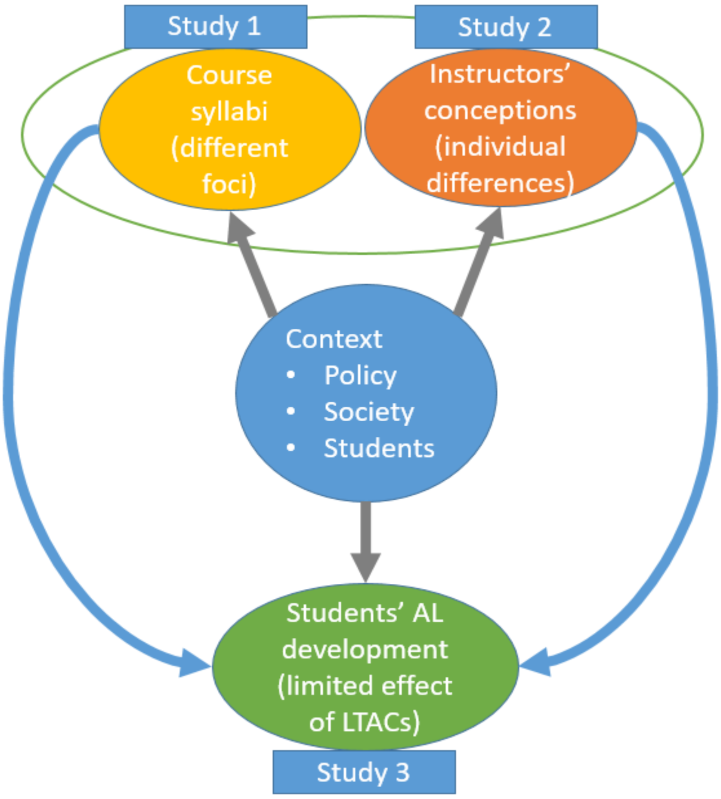
Study 1
Study 2
Course syllabi (different foci)
Instructors' conceptions (individual differences)
Context
• Policy
• Society
• Students
Students' AL development (limited effect of LTACs)
Study 3

Figure 14. Model of influences.

This model identifies the general contextual factors that impinge directly on higher education instructors, course curricula and syllabi and students' assessment literacy development. Additionally, the course instructor and syllabi can be viewed as one unit for analysis which is also intended to impact on students' literacy development. The contextual factors as identified in the thesis include: 1) the policy, i.e., jurisdictional education and assessment policies; 2) the society, i.e., the societal values and norms related to assessment; and 3) the students participating in the courses (the level and conditions of students being taught). These factors co-exist but vary in different contexts and create the overall environment that contributes to the construction and implementation of LTACs and students' assessment literacy development.

Study 1 identified different course foci across the Chinese undergraduate, Chinese postgraduate and NZ postgraduate course syllabi. The major difference lay in the degree to which testing versus assessment more generally was the focus of the course. Chinese undergraduate courses were almost totally about language testing, Chinese postgraduate incorporated more contents and objectives of assessment while New Zealand postgraduate courses presented a more general framework of assessment in which testing was one component. Study 2 interviews with instructors partly echoed the impact of societal context on course design, but revealed that instructors' individual differences in terms of experience and professional learning of assessment made a difference to the courses, most especially visible at the postgraduate level in China (further details in section 7.1.3). Study 3 revealed that the effect of a single assessment preparation course was limited in enhancing undergraduate student teachers' assessment literacy; reflecting the power of various contextual factors (further discussed in section 7.1.4).

In general, this model concurs with previous research (Carless, 2011; Kennedy, 2016), especially Brown and Harris (2009) who identified the impact of contextual factors on teacher actions and beliefs:

> The conceptions of both teachers and students are influenced by various policy directions and priorities and these beliefs, in turn, guide their separate teaching and learning practices. These two pathways are shaped by and respond to societal and cultural contexts, meaning that there will be different beliefs and practise in differing societal, ethnic, and cultural groups. (p.70).

In other words, this thesis found that the predominant assessment cultures of two different nations coloured and controlled how assessment was taught. In China the dominant model of assessment is summative testing, while in New Zealand there is a greater focus on formative assessment which has a much broader scope than summative testing. This study reveals this influence in how courses were designed in the two different cultures. Unsurprisingly, this thesis also shows that the beliefs of student teachers in teacher education programmes about assessment reflect the dominant uses and practices of the system for which they are being trained (e.g., Chen & Brown, 2013). Furthermore, the limited effect of attending an LTA course on student teachers' assessment literacy corroborates some previous research (e.g., Beziat & Coleman, 2015; Brown, 2008; Deneen & Brown, 2016) with practicing teachers, but not others e.g. Mertler (2009) and McGee & Colby (2014). An important contribution of this thesis is the clear identification of the power that individual course instructors have in higher education settings to influence, despite societal norms, what is taught and how it is taught.

7.1.2 The contextual factors.

The policies in different societal contexts work together to form the macro environment that shape the conceptions and practices of both teachers and students. Within the Chinese context, although the national education policy advocates using multiple forms of assessment in an effort to diminish the negative washback effects of testing throughout all education sectors, the existence of highly selective and high-stakes national entrance exams (*zhongkao/gaokao*) is the deciding factor on what future teachers are taught. Chinese society relies on entrance exams to select high-achievers for key secondary schools and universities. A decent job and a stable life in urban areas are more likely for those who get good grades (see section 2.2.1 Testing and its washback effects., Chapter 2). The importance China places on high academic achievement measured by tests has a lasting and profound effect on the education system. Thus the entrance exams become the de facto policy for education and the greatest concern for schools, teachers, students and parents. More importantly, it appears that this summative test-based approach to assessment defines not only what schools teach students but also what teacher educators teach undergraduate prospective teachers.

Kennedy, Chan and Fok (2011) applied the concepts of “soft policy” and “hard policy” to describe the relationship between the explicit national policy toward a broad view of assessment and the implicit ideology that examinations are the only valid assessment that governs teachers and students’ practices. China’s national policy to use multiple forms of assessment lacks power for implementation and functions as a soft policy. Whereas entrance exams give tests and exams supreme power, acting as a hard policy that drives the conceptions and practices of assessment for schools, teachers, and students (Yu & Suen, 2005).

The impact of the “hard policy” and “soft policy” in China is visible in this thesis. The “hard policy” has a restricting effect on course curriculum and instructors in China, resulting in an emphasis on the teaching of testing rather than assessment with a broader perspective. As demonstrated in Study 1 syllabi analysis, Chinese undergraduate courses focused heavily on testing. At the Chinese postgraduate level, there was greater attention on assessment in general, but testing was still the main focus. Study 2 found that Chinese instructors held passive and hesitant attitudes towards national policy and curriculum guidance. The national policy for implementing formative and multiple assessment was perceived as a symbolic policy while the testing-oriented and exam-driven educational environment gave prominence to the "testing" focus at both undergraduate and postgraduate levels.

Study 2 explained this focus: Chinese instructors took the testing-oriented assessment environment into consideration, contextualized their courses by focusing on testing knowledge and skills so that students could be prepared for their future work teaching in Chinese schools. Indeed, seven of the 16 Chinese instructors taught courses called Language Testing and included no content on formative or classroom-based assessment. These instructors believed testing, as a tradition, was the major and sometimes the only way of assessing students. Correspondingly, these instructors' definition of assessment literacy was restricted to testing literacy. On the whole, the Chinese instructors viewed testing and assessment as separate entities, with testing being the normative tradition while assessment was newly emerging.

Within the New Zealand context, the national assessment policy specifies a deliberate focus on assessment for learning principles rather than a narrow testing regime (NZ Ministry of Education, 2007). Both the policy and the education system encourage teachers to assess against national curriculum standards to a great extent through formative assessment while students are encouraged to assess their own learning (Absolum et al., 2009; Crooks, 2002). There is no national nor regional standardised testing before year 10 and New Zealand's university entrance entry is obtained through a mixture of school-based assessments and summative examinations (see section 2.2.3 Assessment policy and environment in New Zealand., Chapter 2). Just as seen in China, instructors' curriculum and instruction were aligned with NZ policy and societal priorities. As identified in Study 1, NZ postgraduate course syllabi used assessment as the umbrella term covering both testing and other forms of assessment. Study 2 corroborated Study 1 findings that 1) NZ instructors conceived of assessment as a comprehensive concept including both testing and other forms of assessment and 2) their definition of assessment literacy was broader with a stronger focus on formative classroom assessment than the Chinese instructors. In addition, NZ instructors' attitudes towards national policy and curriculum guidance was more welcoming and attentive. One of them incorporated NZ language policy as part of the course content while two of them taught students to assess against national assessment principles and standards.

In addition to society and policy, students were also a contextual factor that functioned in the model. As found in Study 2, both Chinese and NZ instructors took students into consideration when designing their courses. Chinese instructors took student teachers' future teaching and learning experiences into consideration, which was mostly associated with China's testing environment. This also explains why there was an overall focus on testing in the Chinese LTACs. NZ postgraduate course instructors took students' multi-cultural and multi-lingual backgrounds into consideration and set their courses in a wider social-cultural and educational environment. Considering the large number of international students (mostly Asian) who were educated in testing cultures, NZ instructors asked students to reflect on their learning experiences and think how to apply the assessment knowledge they learned in NZ in their home countries. Simultaneously, the effect of policy and society on students was obvious. In Study 3, there was little difference around the assessment literacy components for Chinese undergraduate student teachers between those who had or did not have a previous LTA course (discussed further in section 7.1.4). Students were very much concerned about assessment errors and accuracy, perhaps because of the continuing determinative role of the high-stakes entrance exams that govern schooling.

7.1.3 Instructor's individual differences.

Additionally, this thesis has identified the influence of instructors' individual own beliefs, values or attitudes in devising curriculum in higher education. That is to say, the

LTAC curriculum was not just a reflection of jurisdictional policies and societal influences—it is also a product of individual differences in instructors. Instructors' own beliefs, values, and experiences (e.g., professional development in assessment either individual or in group, overseas assessment experiences, participation in assessment research) contributed to the focus of the courses. Study 1 found that the portion of assessment-related objectives and contents in Chinese postgraduate course syllabi was greater than in undergraduate course syllabi. Study 2 further revealed that some Chinese instructors were pushing for a change: Five out of ten Chinese undergraduate instructors and four out of six postgraduate instructors incorporated topics of assessment in their teaching and argued the importance of learning about formative classroom assessment. These Chinese instructors responded to the testing context differently than those with a greater focus on testing in that they tried to break the boundaries of testing and establish new concepts. Their conceptual framework of assessment included both testing and assessment; their definitions of assessment literacy were similar to their NZ counterparts.

The reasons for these individual differences as identified in Study 2 included: instructors' individual learning; institutional community learning; participation in domestic and international academic conferences; and overseas assessment learning and research experiences. This insight is consistent with Ajzen's (1991; 2002; 2005) theory of planned or reasoned behaviour which draws attention to perceived and actual levels of control in explaining behaviour. Ajzen's theory states that one's attitudes towards behaviour, how they think others want them to perform the behaviour (subjective norm) and how much they think they can control the behavior (perceived behavior control), shape one's intentions and behaviours. As seen from the study, the Chinese course instructors who pushed for a change to incorporate assessment into their courses regarded formative assessment as effective assessment methods to 1) counterbalance the negative washback effects of testing and 2) improve teaching and learning. They were able to implement their own understanding because of the freedom given to tertiary teachers to exercise control of their own course curricula and teaching. The instructors' differences led to different course construction and implementation, which may also follow from the fact that higher education permits instructors freedom to draw on their individual expertise to design and implement curriculum. In the compulsory school sectors in China there is much less freedom in this respect and greater control on individuals (Wang & Lam, 2009). Compared with teaching in primary and secondary sectors, where assessment is oriented by and centered on exams, it is more plausible for teachers in the tertiary sector to carry out a variety of assessment activities in and outside the classroom where students no longer have to face such pressure and when achievement is no longer evaluated by the marks of students. Thus, in the context of real control it was easier for individuals to implement formative assessment within the policy of testing.

Perhaps, also because of the greater qualifications and exposure to international approaches, the Chinese postgraduate instructors were able to implement a curriculum that took seriously the official policy of reducing emphasis on testing. This level of personal expertise is mimicked in the New Zealand postgraduate instructors who were all PhD holders, regular attenders at international conferences, and contributing publishers in the international literature. Where instructors have greater expertise, they are more able to exercise autonomous control over their teaching.

7.1.4 The limited course effects.

The formation of students' assessment literacy can be depicted as a twisted spiral of many cords: two of which are their previous learning and assessment experiences as a student and the cognitive process through their teacher education programme, which are all shaped by the external, the social cultural environment and practicum (Eyers, 2014; Hill et al. 2011). The primary goal for Chinese assessment preparation courses, as per Study 1 and 2, is to develop students' capacity to function in the world of school testing. Consequently, Study 3 discovered that there were limited effects of LTACs on Chinese undergraduate student teachers' assessment literacy, namely, conceptions of assessment, self-efficacy for assessment, and self-reported practices of assessment. Multi-group invariance testing showed the structural relationship between the three components were all equivalent between LTAC-participants and non-participants, suggesting the experience of a course did not change how students understood assessment as testing. There are several possible reasons that contribute to this limited impact. Firstly, as found in Study 1 and Study 2, the attention paid to assessment preparation rather than testing preparation was insufficient in the Chinese undergraduate context. Not every student was required to take such a course and when they were required they tended to be relatively short (i.e., 8 to 16 weeks). It is unlikely a single short course focused on language testing is powerful enough to enhance students' general assessment literacy. Second, consciously or unconsciously, teachers learn how to teach through their learning experience as a student in school (Borg, 2015; Pajares, 1992). The pre-service teachers' assessment experiences before their undergraduate study were mostly testing and preparing for the high-stakes national entrance exams, which has strongly shaped their conceptual model of assessment as testing and can be too deep-rooted to change (Green, 1971). Thus, as identified in Study 3, students were very concerned about the measurement errors in assessment. Even though some Chinese instructors include more formative approaches into their teaching, the societal context exert a strong influence (as reviewed in section 5.4, Chapter 5). Thirdly, according to Zhan (2008), China's pre-service teacher education focused too much on subject content knowledge but neglect the pedagogy and assessment training. These all contribute to the limited impact of LTACs on students' assessment literacy development.

However, the factor mean score comparison in Study 3 indicated LTAC participants had higher level of self-efficacy; higher acceptance of assessment as helping learning, taking errors into consideration in assessment and higher level of three forms of assessment practices. These results indicated that assessment preparation courses had a small positive effect in boosting self-efficacy level for assessment with a more comprehensive view on assessment.

7.2 Implications

Overall, this thesis advances our understanding about assessment preparation courses and assessment literacy development in the contexts of China and NZ. It raises theoretical and practical implications for other contexts where policy and practice conditions are similar.

7.2.1 Theoretical implications

There are several theoretical implications stemming from this thesis. To begin with, the thesis sheds light on the borrowing of the Western idea of “formative assessment” in testing-cultures like China. The concept of formative assessment first emerged as an alternative to the traditional summative assessment of testing first in Western education systems. With the global educational paradigm shift from summative assessment to formative, its implementation in the Western countries where testing was not prevalent seemed to be relatively easy. In contrast, formative assessment in testing cultures like China encountered numerous challenges caused by the complicated teaching environment and deeply-rooted cultural traditions (Carless, 2011; Lam, 2016; Poole, 2016). The Chinese instructors’ conceptual framework of assessment (refer to *Figure 8* in section 5.4.1, Chapter 5) reflects the development of Chinese instructors’ conceptions of assessment. Assessment started to emerge as a separate circle alongside testing. With the development and translation of formative assessment in China, it is highly likely that the circle of assessment will continue to enlarge. However, whether it will become an inclusive circle that incorporates testing (like the NZ instructors’ figure) is doubtful, despite a policy framework that permits it. Unless there are radical reforms in the Chinese education system, it is highly likely that the hegemonic practice of testing and exam preparation will prevail and if assessment is incorporated at all it will only be used to support this entrenched practice.

Second, as Study 3 revealed, Chinese undergraduate pre-service teachers’ self -efficacy of assessment functioned as a mediator between conceptions and practices of assessment. The role of self-efficacy as a predictor of behaviour is established, which enhanced the field of teacher conceptions and adds to previous research on the relationship between conceptions and practices of assessment. Theoretically, assessment practices could be enhanced when conceptions and self-efficacy for assessment were both improved.

7.2.2 Practical implications.

Based on the findings and discussions as well as literature review (e.g., Brown, 2008; Denee & Brown, 2016), one single assessment preparation course is not enough for student teachers' assessment literacy development; rather, it should be a systematic and on-going process that integrates assessment courses, teacher education programmes, student teacher practicum and teacher professional development within a culture that values and promotes ideas and values of comprehensive assessment (Deluca et al., 2013; Taylor, 2009; Volante & Fazio, 2007). Thus, the practical implications are proposed at the following three levels: for assessment course instructors; for teacher education programme leaders; and for the Ministry of Education officials.

7.2.2.1 Assessment course instructors.

One of the most fundamental aspects for course instructors in assessment preparation course is to know what assessment entails (Berger, 2012). As seen from this thesis, for instructors from testing cultures such as China, it is important to take account of the contextual factors including the social-cultural educational environment and the students' pre-existing conceptions to strike a balance between testing and formative classroom assessment. Teachers should be conceptually aware of the wide variety of assessment practices such as, teacher observation, classroom interaction, portfolios and paper-based exams. This is necessary in order to help students develop a more comprehensive understanding and application of assessment. As pointed out in section 2.1.2.4 The relationship between SA and FA., Chapter 2, the distinction between summative and formative assessment can be superficial. Instructors should know that there is a wide range of assessment tools beyond paper-based standardised tests and assessment serves multiple purposes from selection, improvement of learning, controlling and accountability. Besides, it is suggested that Chinese instructors incorporate national policy in their assessment courses to inform students of national policy and curriculum guidance of using various assessment methods. The NZ instructors took students' home culture and learning experiences into consideration in course construction in a practice-based approach. This sets an example with implications for instructors in non-testing cultures who educate international students from various backgrounds. It is essential to consider students' previous assessment conceptions and experiences shaped by their home culture. Using self-reflection and comparison as a tool could help student teachers conceptualise and apply the knowledge principles and skills of assessment so as to better prepare them for the possible teaching contexts where they would teach.

In addition, the practical application of language assessment is an important component in assessment literacy development. This is not only identified in the literature (Boyles, 2006; Inbar-Lourie, 2008; Stiggins, 1995) but also evidenced in the thesis findings. Thus, it is suggested for instructors in all contexts to provide as many opportunities for practice as possible for students to develop their assessment literacy.

7.2.2.2 Teacher education programme leaders.

Considering the short time span and limited effect of one single assessment course, it is suggested that assessment preparation courses should be made as a compulsory part of teacher education programmes in all contexts. Furthermore, at least two courses on assessment could be considered, one to address issues on testing while the other on formative classroom assessment. It's also recommended that assessment literacy preparation be integrated within the overall teacher education programme. Assessment literacy development should be the responsibility of all teachers, instructors and leaders involved in the teacher education programme rather than the sole responsibility of one single course instructor. Teaching should be done with multiple assessments throughout other courses to increase the exposure of student teachers to assessment practices. When all the aspects of teaching and assessment are tuned to support comprehensive understanding of assessment, students' learning about assessment can be optimised. Curriculum-based assessment composed of tests and classroom-based assessment activities can be designed and implemented to integrate policy, curriculum, teaching and assessment in a systematic whole to ensure the correspondence between tests and policy. These implications apply to both China and NZ contexts.

Another aspect to consider is the role of practicum in China's undergraduate pre-service teacher education programme. As evidenced from previous research (Smith et al., 2014), practicum serves as an important role in student teachers' assessment literacy development. Practicum provides authentic teaching environment for student teachers to practice and reflect on the assessment knowledge and skills taught in the programme. However, most Chinese undergraduate teacher education programmes only organise one practicum during the entire teacher education programmes (Yan & He, 2010). Considering the insufficiency of practicum in China's teacher education programme, it is suggested practicum be carried out once every year. Furthermore, to ensure sound assessment practice be implemented, an associate teacher with excellent assessment literacy should accompany student teachers during practicum to guide their practice.

7.2.2.3 Ministry of Education officials.

At the policy level, considering the conflicts between the "hard policy" and the "soft policy" in China, the Chinese Ministry of Education should take measures to ensure that the assessment policy advocating multiple forms of assessment is implemented and supported at various levels. As implied from Study 2, professional learning in assessment led to changes of individual instructors' conceptions. Research has also confirmed that teachers involved in ongoing professional development in assessment significantly increase their assessment literacy level and gained better understanding of authentic assessment (e.g., Koh, 2011; Lieberman, 1995; Light, Calkins & Cox, 2009). The Chinese Ministry of Education should provide tertiary instructors with opportunities and resources for professional learning such as attending domestic and international academic conferences for intellectual exchange and cooperation, gaining higher degrees and encouraging publications. This is important to update instructors' knowledge and skills to enhance their conceptions and thus to adjust practices that fit in the contexts and various requirements. However, it should be brought to attention that teachers' ability and power of using formative assessment practices is largely constrained in text-centric curricula and education system (Deneen et al., 2019; Zhao et al., 2018). From the micro everyday classroom teaching to the macro external education system, a balanced assessment scheme combining both summative testing and formative and classroom assessment should be designed and implemented to support formative assessment. A more daunting implication is to remove the entrance exams in the Chinese education system.

7.3 Limitations and suggestions for further research

Due to time and resource limits not all important issues were addressed. Classroom observations with LTAC course instructors were not carried out. Future research that observes instructors' assessment practices would add evidence as to how LTACs courses were actually taught in the classroom.

In the student teachers' survey, the SEM makes a causal claim about the relationship between self-efficacy, conceptions, and practices of assessment using correlational data. This points to a causal relationship that needs to be tested in future intervention studies in which input variables are manipulated to ascertain whether the observed patterns in this study lead to real-world changes. In additon, only the effect of LTACs has been examined, it would be insightful to examine the effect of practicum on assessment literacy development as suggested in previous studies of general teacher education students' assessment learning in NZ and other contexts. While it is highly likely that practicum functions as a strong and positive predictor for student teachers' assessment literacy development, it is unknown whether students actually engage in assessment as part of standard practicum experience. There is scope for further research in this area.

Another limitation is that this thesis only looked at student teachers in the Chinese undergraduate context and thus there was no student teacher data from Chinese postgraduate nor NZ postgraduate contexts. It would be enlightening to investigate postgraduate students' responses to the survey, so as to better establish whether courses with a stronger emphasis on assessment produce teachers with a more encompassing understanding of testing and assessment. Another way to build on this thesis would be to examine how, or if, teachers' assessment literacy components change after they enter into a teaching career. This would ideally be done across disparate educational, social and cultural environments to ascertain how powerful contextual factors are. This would also raise further questions about what it takes to change conceptions and practices in a schooling system of various contexts.

Considering the unchanged testing-oriented assessment environment in China, it would be interesting and meaningful to conduct a longitudinal study on pre-service teachers' professional development trajectory in terms of how their conceptions, self-efficacy and practice of assessment changed, if at all, when they step into their career. Also, given the increasing preparation and employment of teachers from China in the NZ context to teach Mandarin and ESOL, it may also be useful to understand the effects of Chinese cultural heritage on teachers teaching Mandarin Chinese in NZ who undertake their teacher preparation in and teach in NZ context.

7.4 Conclusion

Over the past decades, language teacher education has appealed for more research and attention on innovative approaches to assessment and assessment literacy preparation and development. This thesis explored how LTACs were constructed and implemented in three cultural and academic contexts with an intention of exploring how contextual factors influence LTACs and how LTACs help develop student teachers' assessment literacy. Based on a pragmatist paradigm, an exploratory sequential mixed-methods research design was employed to explore three aspects of LTACs (course syllabi, course instructors' interviews, and students' survey response). Three studies were conducted and linked to form a data triangulation, which deepened our understanding of the nature of assessment literacy preparation courses in different cultural and academic contexts.

The findings and discussions identified a model of influences on the assessment preparation courses and assessment literacy preparation. This thesis also reinforced the notion that teacher assessment literacy development, like any other area of development, involves many factors. In addition to the influence of external societal factors, how teachers develop their assessment literacy can be regarded as an embedded activity shaped by the internal local context with teachers, students and the teaching environment. It is a combined result of one's prior learning experience as a student, teacher preparation and professional development, along with the teaching environment s/he is in.

The original and substantive contribution for this thesis includes:

1) The insight that instructors' individual conceptions make a difference to curriculum construction and implementation in higher education context. These changes of their conceptions are due to their research experiences, their professional development in assessment and their international learning and research conference experiences.
2) This thesis also highlights possibilities for the future: it is expected that with the increasingly frequent intellectual exchanges between nations, the Chinese instructors will bridge the gap with their international counterparts as holding a more comprehensive understanding of assessment while taking the external context into consideration. On the other hand, assessment course instructors in the western tertiary context like NZ will take the testing-culture into consideration and cater to the needs of the international students.
3) As evidenced in the Chinese undergraduate pre-service teacher survey, conceptions, self-efficacy and practices of assessment are three important components for teacher assessment literacy development, in which self-efficacy functions as a mediator between conceptions and practices. This adds to previous studies on relationship between teacher beliefs and practices by inserting self-efficacy as a mediator between models of teachers' conceptions and practices of assessment.

References

Absolum, M., Flockton, L., Hattie, J., Hipkins, R., & Reid, I. (2009). *Directions for assessment in New Zealand: Developing students' assessment capabilities.* Wellington, New Zealand: Ministry of Education.

Airasian, P. W. (1991). Perspectives on measurement instruction. *Educational Measurement: Issues and Practice, 10*(1), 13.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Ajzen, I. (2002). Residual effects of past on later behavior: Habituation and reasoned action perspectives. *Personality and Social Psychology Review, 6*(2), 107-122.

Ajzen, I. (2005). *Attitudes, personality and behavior* (2nd ed.). New York, NY: Open University Press.

Alkharusi, H. (2011a). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia-Social and Behavioral Sciences*, *29*, 1614-1624.

Alkharusi, H. (2011b). A logistic regression model predicting assessment literacy among in-service teachers. *Journal of Theory and Practice in Education*, *7*(2), 280-291.

Alkharusi, H., Kazem, A. M., & Al-Musawai, A. (2011). Knowledge, skills, and attitudes of preservice and in-service teachers in educational measurement. *Asia-Pacific Journal of Teacher Education*, *39*(2), 113-123.

American Federation of Teachers, National Council on Measurement in Education, & National Education Association (AFT/NCME/NEA). (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and Practice, 9*(4), 30-32.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives*, *10*, 18. Retrieved Dec. 2017 from http://epaa.asu.edu/epaa/v10n18/.

Angelo, T. A., & Cross, K. P. (2012). *Classroom Assessment Techniques*. San Francisco, CA: Jossey Bass Wiley.

Archer, E., & Brown, G. T. L. (2013). Beyond rhetoric: Leveraging learning from New Zealand's assessment tools for teaching and learning for South Africa. *Education as Change*, *17*(1), 131-147.

Atkins, L., & Wallace, S. (2012). *Qualitative Research in education.* New York, NY: Sage.

Aveyard, H. (2014). *Doing a literature review in health and social care: A practical guide.* Berkshire, UK: McGraw-Hill Education.

Bailey, K. M., & Brown, J, D. (1998). Language testing courses: what are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 236-258). Great Britain: Cromwell Press.

Bailey, K., & Curtis, A. (2015). *Learning about language assessment: Dilemmas, decisions, and directions*. United States: National Geographic Learning.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bandalos, D. L., & Finney, S. J. (2010). Factor analysis: Exploratory and confirmatory. In G. R. Hancock & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences* (pp. 93-114). New York, NY: Routledge.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Bandura, A. (2010). Self- efficacy. In I.B. Weiner and W.E. Craighead (Eds.), *The Corsini Encyclopedia of Psychology*, pp (1-3). doi:10.1002/9780470479216.corpsy0836.

Banta, T. W. (Ed.). (2002). *Building a scholarship of assessment*. San Francisco, CA: John Wiley & Sons.

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual differences*, *42*(5), 815-824.

Beavers, A., Lounsbury, J., Richards, J., Hush, S., Skolits, G., & Esquivel, S. (2013). Practical considerations for using exploratory factor analysis in educational research. *Practical Assessment, Research and Evaluation*, 18, 1–13.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*(2), 70-91.

Bentler, P. M. (1978). The interdependence of theory, methodology, and empirical data: Causal modeling as an approach to construct validation. In D. B. Kandel (Ed.), *Longitudinal Drug Research* (pp. 267-302). New York, NY: John Wiley & Sons.

Berger, A. (2012). Creating Language-Assessment Literacy: A Model for Teacher Education. *Theory and practice in EFL teacher education*, *3*(5), 57-82.

Berry, R. (2008). *Assessment for learning* (Vol. 1). Hong Kong: Hong Kong University Press.

Berry, R. (2011). Assessment trends in Hong Kong: Seeking to establish formative assessment in an examination culture. *Assessment in Education: Principles, Policy & Practice*, *18*(2), 199-211.

Berry, R., & Adamson, B. (Eds.). (2011). *Assessment reform in education: policy and practice* (Vol. 14). New York, NZ: Springer Science & Business Media.

Berry, V., & O'Sullivan, B. (2016, September). *Assessment literacy for language teachers.* Paper presented at the IATEFL 2016: The International Association of Teachers of English as a Foreign Language, Birmingham.

Beziat, T. L., & Coleman, B. K. (2015). Classroom assessment literacy: Evaluating pre-service teachers. *The Researcher*, *27*(1), 25-30.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher education*, *32*(3), 347-364.

Biggs, J. (1998). Assessment and classroom learning: a role for summative assessment. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 103-110.

Biggs, J. (2003). Aligning teaching and assessment to curriculum objectives. *Imaginative Curriculum Project,* LTSN Generic Centre. Retrieved April, 2018, from https://www.advance-he.ac.uk/knowledge-hub/aligning-teaching-and-assessment-curriculum-objectives.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215-232.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8-21.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, *5*(1), 7-74.

Black, P., & Wiliam, D. (2006). Assessment for learning in the classroom. In J. Gardner (Ed.), *Assessment and Learning* (pp. 9-25). London: SAGE Publications.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 1(1), 5-31.

Bloom, B. S., Hastings, J. T., and Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*. New York, NY: McGraw-Hill.

Boaler, J., & Humphreys, C. (2005). *Connecting mathematical ideas: Middle school video cases to support teaching and learning*. Portsmouth, NH: Heinemann.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models* (Vol. 154). Newbury Park, CA: Sage.

Booth, D. (2012). Scaffolding during the formal assessment of young EAL learners. *Australian Review of Applied Linguistics*, *35*(1), 5-27.

Borg, S. (2015). *Teacher cognition and language education: Research and practice*. New York, NY: Bloomsbury Academic.

Boyles, P. (2006). Assessment literacy. In Rosenbusch M. H. (Eds), *New Visons in Action: National Assessment Summit Papers* (pp. 11-15). Iowa City, Iowa: Iowa State University.

Bray, M., Adamson, B., & Mason, M. (2014). *Comparative education research: Approaches and methods* (Vol. 19). Singapore: Springer.

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

Braun, V., & Clarke, V. (2013). *Successful Qualitative Research a practical guide for researchers: A practical guide for beginners*. London, UK: Sage.

Brindley, G. (2001). Language assessment and professional development. In Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T. & O'Loughlin, K. (Eds.), *Experimenting with Uncertainty: Essays in Honor of Alan Davies, 11*, 137-143.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational Measurement: Issues and Practice*, *22*(4), 5-12.

Brown, G. T. (2002). *Teachers' Conceptions of Assessment*. (Doctoral thesis, University of Auckland, Auckland, New Zealand).

Brown, G. T. (2004). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological reports*, *94*(3), 1015-1024.

Brown, G. T. (2008). Assessment literacy training and teachers' conceptions of assessment. In Rubie-Davies, C. M. & Rawlinson C. (Eds.), *Challenging Thinking about Teaching and Learning* (pp. 269-285). New York, NY: Nova Science Publishers.

Brown, G. T. (2011). New Zealand prospective teacher conceptions of assessment and academic performance: Neither student nor practicing teacher. In Kahn R., McDermott J. C., Akimjak A. (Eds.), *Democratic Access to Education* (pp.119-132). Los Angelos, CA: Antioch University.

Brown, G. T. (2018). *Assessment of Student Achievement.* London, UK: Routledge.

Brown, G. T. (2019). Is assessment for learning really Assessment?. In *Frontiers in Education* (Vol. 4, p. 64). Frontiers. Retrieved Feb, 2017, from https://doi.org/10.3389/feduc.2019.00064.

Brown, G. T., Chaudhry, H., & Dhamija, R. (2015). The impact of an assessment policy upon teachers' self-reported assessment beliefs and practices: A quasi-experimental study of Indian teachers in private schools. *International Journal of Educational Research, 71,* 50-64.

Brown, G. T., & Gao, L. (2015). Chinese teachers' conceptions of assessment for and of learning: Six competing and complementary purposes. *Cogent Education*, *2*(1), 993836.

Brown, G. T., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of Multi-Disciplinary Evaluation*, 6(12), 68-91.

Brown, G. T., Harris, L. R., O'Quin, C., & Lane, K. E. (2017). Using multi-group confirmatory factor analysis to evaluate cross-cultural research: Identifying and understanding non-invariance. *International Journal of Research & Method in Education*, *40*(1), 66-90.

Brown, G. T., Hui, S. K., Flora, W. M., & Kennedy, K. J. (2011). Teachers' conceptions of assessment in Chinese contexts: A tripartite model of accountability, improvement, and irrelevance. *International Journal of Educational Research*, *50*(5-6), 307-320.

Brown, G. T., Irving, S. E., & Keegan, P. J. (2014). *An introduction to educational assessment, measurement & evaluation: Improving the quality of teacher-based assessment (Third ed.)*. Auckland, New Zealand: Dunmore Publishing.

Brown, G. T., Lake, R., & Matters, G. (2011). Queensland teachers’ conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education*, *27*(1), 210-220.

Brown, G. T., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: Understanding Hong Kong teachers’ conceptions and practices of assessment. *Assessment in education: principles, policy & practice*, *16*(3), 347-363.

Brown, G. T., & Remesal, A. (2012). Prospective teachers' conceptions of assessment: A cross-cultural comparison. *The Spanish Journal of Psychology, 15*(1), 75-89.

Brown, H., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007?. *Language Testing*, *25*(3), 349-383.

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, *32*(4), 653-675.

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: A road less traveled. *Structural Equation Modeling*, *11*(2), 272-300.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*(4), 872-882.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. New York, NY: Routledge.

Campbell, J., & Li, M. (2008). Asian students' voices: An empirical study of Asian students' learning experiences at a New Zealand university. *Journal of Studies in International Education*, *12*(4), 375-396.

Carless, D. (2011). *From testing to productive student learning: Implementing formative assessment in Confucian-heritage settings*. New York, NY: Routledge. Retrieved from https://doi.org/10.4324/9780203128213.

Chen, P. P. (2005). Teacher candidates' literacy in assessment. *Academic Exchange Quarterly, 9*(3), 62-67.

Chen, J., & Brown, G. T. L. (2013). High-stakes examination preparation that controls teaching: Chinese prospective teachers' conceptions of excellent teaching and assessment. *Journal of Education for Teaching, 39*(5), 541-556.

Chen, J., & Brown, G. T. (2016). Tensions between knowledge transmission and student-focused teaching approaches to assessment purposes: Helping students improve through transmission. *Teachers and Teaching, 22*(3), 350-367.

Cheng, L. (2005). *Changing language teaching through language testing: A washback study* (Vol. 21). Cambridge, UK: Cambridge University Press.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15-37.

Cheng, L., & Qi, L. (2006). Description and examination of the national matriculation English test. *Language Assessment Quarterly: An International Journal*, *3*(1), 53-70.

Cheng, L., & Curtis, A. (2010). *English language assessment and the Chinese learner*. New York, NY: Routledge.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*(2), 233-255.

Chinese Ministry of Education. (2011). *English curriculum standards for compulsory education* (experimental edition). Beijing: Beijing Normal University Press.

Chinese Ministry of Education. (2001). *English syllabus for nine-year compulsory education in full-time junior and secondary schools*. Beijing: People's Education Press.

Chinese Ministry of Education. (2007). *College English Curriculum Requirements*. Beijing: Foreign Language Teaching and Research Press.

Chinese Ministry of Education. (2011). *English curriculum standards for compulsory education* (2011 edition). Beijing: Beijing Normal University Press.

Colapietro, V. M. (2006). Charles Sanders Pierce. In J. R. Shook & J. Margolis (Eds.), *A Companion to Pragmatism* (pp. 13-29). Malden, MA: Blackwell.

Comeaux, E., Brown, A., & Sieben, N. P. (2015). Issues in athletic administration: A content analysis of syllabi from intercollegiate athletics graduate courses. *Innovative Higher Education*, 40(4), 359-372.

Coombe, C., O'Sullivan, B., & Stoynoff, S. (2012). *The Cambridge guide to second language assessment.* New York, Cambridge University Press.

Cortazzi, M., Pilcher, N., & Jin, L. (2011). Language choices and ‘blind shadows’: investigating interviews with Chinese participants. *Qualitative Research*, *11*(5), 505-535.

Crawford, R. J. (2016). *History of tertiary education reforms in New Zealand*. New Zealand Productivity Commission, Te Kōmihana Whai Hua Aotearoa.

Crawford, S. (2002). Evaluation of web survey data collection systems. *Field Methods*, *14*(3), 307-321.

Creswell, J. W. (2009). *Mapping the field of mixed methods research*. London, UK: Sage.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th ed.). San Francisco, CA: Sage.

Creswell, J. W. (2014). *A Concise Introduction to mixed methods research*. New York, NY: Sage.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. San Francisco, CA: Sage.

Crooks, T. J. (2002). Educational assessment in New Zealand schools. *Assessment in Education: Principles, Policy & Practice, 9*(2), 237-253.

Crooks, T. (2004, March). *Tensions between assessment for learning and assessment for qualifications*. Paper presented at the Third Conference of the Association of Commonwealth Examinations and Accreditation Bodies (ACEAB), Nadi, Fiji.

Crooks, T. J. (2010). New Zealand: Empowering teachers and children. *Balancing change and tradition in global education reform*, 281-310.

Cummins, J., & Davison, C. (Eds.). (2007). *International handbook of English language teaching* (Vol. 15). New York, NY: Springer Science & Business Media.

Damon, W. (2007). Dispositions and teacher assessment: The need for a more rigorous definition. *Journal of Teacher Education*, *58*(5), 365-369.

Davey, G., Lian, C., & Higgins, L. (2007). The university entrance examination system in China. *Journal of further and Higher Education*, *31*(4), 385-396.

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing, 25*(3), 327-347.

De Rada, V. D., Ariño, L. V. C., & Blasco, M. G. (2016). The use of online social networks as a promotional tool for self-administered internet surveys. *Revista Española de Sociología (RES)*, *25*(2), 189-203.

DeLuca, C., Chavez, T., Bellara, A., & Cao, C. (2013). Pedagogies for preservice assessment education: Supporting teacher candidates' assessment literacy development. *The Teacher Educator, 48*(2), 128-142.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates’ learning. *Assessment in Education: Principles, Policy & Practice, 17*(4), 419-438.

DeLuca, C., Klinger, D., Pyper, J., & Woods, J. (2015). Instructional rounds as a professional learning model for systemic implementation of assessment for learning. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 122-139.

DeLuca, C., & Lam, C. Y. (2014). Preparing teachers for assessment within diverse classrooms: An analysis of teacher candidates' conceptualizations. *Teacher Education Quarterly, 41*(3), 3-24.

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Teacher assessment literacy: A review of international standards and measures. *Educational Assessment, Evaluation and Accountability*, *28*(3), 251-272.

Deneen, C. C., & Brown, G. T. (2016). The impact of conceptions of assessment on assessment literacy in a teacher education program. *Cogent Education*, *3*(1), 1225380.

Deneen, C. C., Fulmer, G. W., Brown, G. T. L., Tay, H. W., Tan, K., & Leong, W. S. (2019). Value, practice and proficiency: Teachers' complex relationship with Assessment for Learning. *Teaching and Teacher Education*, 80, 39-47. doi:10.1016/j.tate.2018.12.022.

Dennen, V. P., Aubteen D., A., & Smith, L. J. (2007). Instructor–learner interaction in online courses: The relative perceived importance of particular instructor actions on performance and satisfaction. *Distance education*, *28*(1), 65-79.

DeVellis, R. (2003). *Scale development: theory and applications: theory and application*. Thousand Okas, CA: Sage.

Denzin, N. K. (2008). *Collecting and interpreting qualitative materials* (Vol. 3). New York, NY: Sage.

Dixon, J. K., Andreasen, J. B., & Stephan, M. (2009). Establishing social and sociomathematical norms in an undergraduate mathematics content course for prospective teachers: The role of the instructor. *Scholarly Practices and Inquiry in the Preparation of Mathematics Teachers,* 43.

Dunlap, J. C. (2005). Problem-based learning and self-efficacy: How a capstone course prepares students for a profession. *Educational Technology Research and Development, 53*(1), 65-83.

Earl, L. M. (2012). *Assessment as learning: Using classroom assessment to maximize student learning.* San Francisco, California: Corwin Press.

East, M. (2016). *Assessing foreign language students' spoken proficiency: Stakeholder perspectives on assessment innovation* (Vol. 26). Singapore: Springer.

Edwards, F. (2017). A rubric to track the development of secondary pre-service and novice teachers' summative assessment literacy. *Assessment in Education: Principles, Policy & Practice, 24*(2), 205-227.

Edwards, R., & Holland, J. (2013). *What is qualitative interviewing?*. London, UK: Bloomsbury Academic.

Education New Zealand. (2018, February 28). *Latest international student enrollment and student visa trends*. Retrieved May, 2018, from https://enz.govt.nz/assets/Uploads/Latest-international-student-enrolment-and-student-visa-trends2.pdf

Entwistle, N. (2000, June). *Promoting deep learning through teaching and assessment: conceptual frameworks and educational contexts*. Paper presented at the First Annual Conference ESRC Teaching and Learning Research Programme, University of Leicester, U.K.

Etikan, I., Musa, S. A., & Alkassim, R. S. (2016). Comparison of convenience sampling and purposive sampling. *American Journal of Theoretical and Applied Statistics, 5*(1), 1-4.

Eyers, G. (2014). *Preservice teachers' assessment learning: Change, development and growth* (Doctoral dissertation). University of Auckland, Auckland, New Zealand.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4*(3), 272.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509-529.

Field, A. (2016). *An adventure in statistics: The reality enigma*. New York, NY: Sage.

Feilzer, Y. M. (2010). Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm. *Journal of Mixed Methods Research, 4*(1), 6-16.

Fleming, C. M., & Bowden, M. (2009). Web-based surveys as an alternative to traditional mail methods. *Journal of Environmental Management, 90*(1), 284-292.

Foldnes, N., Foss, T., & Olsson, U. H. (2012). Residuals and the residual-based statistic for testing goodness of fit of structural equation models. *Journal of Educational and Behavioral Statistics, 37*(3), 367-386.

Franken, M. P., & McComish, J. (2003). *Improving English language outcomes for students receiving ESOL services in New Zealand schools, with a particular focus on new immigrants*. Wellington: New Zealand Ministry of Education.

Freeman, D., & Johnson, K. E. (1998). Reconceptualizing the knowledge- base of language teacher education. *TESOL Quarterly*, *32*(3), 397-417.

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113-132.

Gebril, A., & Brown, G. T. (2014). The effect of high-stakes examination systems on teacher beliefs: Egyptian teachers' conceptions of assessment. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 16-33.

Gipps, C. (2011). *Beyond Testing (Classic Edition): Towards a theory of educational assessment*. Abington, Oxon: Routledge.

Graham, P. (2005). Classroom-based assessment: Changing knowledge and practice through preservice teacher education. *Teaching and Teacher Education, 21*(6), 607-621.

Graves, R., Hyland, T., & Samuels, B.M. (2010). Understanding writing assignment: An analysis of syllabi at one Canadian college. *Written Communication, 27(3)*, 293-317.

Green, S., & Johnson, R. L. (2010). *Assessment is essential (1st ed.)*. Boston, MA: McGraw-Hill Higher Education.

Green, T. F. (1971). *The activities of teaching*. New York, NY: McGraw-Hill.

Griffiths, C., & Parr, J. M. (2001). Language-learning strategies: Theory and perception. *ELT Journal*, *55*(3), 247-254.

Gu, P. Y. (2014). The unbearable lightness of the curriculum: what drives the assessment practices of a teacher of English as a Foreign Language in a Chinese secondary school? *Assessment in Education: Principles, Policy & Practice, 21*(3), 286-305.

Gu, Y. (2016, July). *Classroom assessment literacy for EFL teachers: Conceptualisation, operationalisation, and validation*. Presented at the LTRC 2016: The 40th Language Testing Research Colloquium, Auckland, New Zealand.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of Qualitative Research*, *2*, 105-117.

Guo, Y., Connor, C. M., Yang, Y., Roehrig, A. D., & Morrison, F. J. (2012). The effects of teacher qualification, teacher self-efficacy, and classroom practices on fifth graders' literacy outcomes. *The Elementary School Journal*, 113(1), 3-24.

Haertel, G. D., Walberg, H. J., & Haertel, E. H. (1991). Socio- psychological environments and learning: A quantitative synthesis. *British Educational Research Journal, 7*(1), 27-36.

Hailaya, W., Alagumalai, S., & Ben, F. (2014). Examining the utility of Assessment Literacy Inventory and its portability to education systems in the Asia Pacific region. *Australian Journal of Education*, *58*(3), 297-317.

Han, M., & Yang, X., (2010). Educational assessment in China: Lessons from history and future prospects. *Assessment in Education: Principles, Policy & Practice, 8*(1), 5-10.

Harlen, W. (2005a). Teachers' summative practices and assessment for learning: tensions and synergies. *Curriculum Journal*, *16*(2), 207-223.

Harlen, W. (2005b). Teaching, learning and assessing science 5-12. London, UK: Sage.

Harlen, W. (2007). *Assessment of learning*. London, UK: Sage.

Harlen, W., & Deakin, C. R. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice*, *10*(2), 169-207.

Harlen, W., & James, M. (1997). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education, 4*(3), 365-379.

Harzing, A. W., Reiche, B., & Pudelko, M. (2013). Challenges in international survey research: A review with illustrations and suggested solutions for best practice. *European Journal of International Management*, *7*(1), 112-134.

Hattie, J. A., & Brown, G. T. (2007). Technology for school-based assessment and assessment for learning: Development principles from New Zealand. *Journal of Educational Technology Systems, 36*(2), 189-201.

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

Hill, M., Cowie, B., Gilmore, A., & Smith, L. F. (2010). Preparing assessment-capable teachers: What should preservice teachers know and be able to do. *Assessment Matters*, *2*, 43-64.

Hirschfeld, G. H., & Brown, G. T. (2009). Students' conceptions of assessment: Factorial and structural invariance of the SCoA across sex, age, and ethnicity. *European Journal of Psychological Assessment*, *25*(1), 30-38.

Hooper, D., Coughlan, J., & Mullen, M. (2008a). Structural equation modelling: Guidelines for determining model fit. *Journal of Business Research Methods*, 6, 53–60.

Hooper, D., Coughlan, J., & Mullen, M. (2008b, September). Evaluating model fit: a synthesis of the structural equation modelling literature. In *7th European Conference on research methodology for business and management studies* (pp. 195-200).

Houston, D., & Thompson, J. N. (2017). Blending formative and summative assessment in a capstone subject: ‘it’s not your tools, it’s how you use them’. *Journal of University Teaching & Learning Practice*, *14*(3), 2-27.

Hoyle, R. H. (1995). *Structural equation modeling: Concepts, issues, and applications*. Thousand Oaks, California: Sage.

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277-1288.

Hu, G. (2003). English language teaching in China: Regional differences and contributing factors. *Journal of Multilingual and Multicultural Development*, *24*(4), 290-318.

Hu, G. (2005). English language education in China: Policies, progress, and problems. *Language Policy*, *4*(1), 5-24.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385-402.

Jian, H., & Luo, S. (2014). Formative assessment in L2 classroom in China: The current situation, predicament and future. *Indonesian Journal of Applied Linguistics, 3*(2), 18-34.

Jiang, H. (2015). *Learning to teach with assessment: A student teaching experience in China.* Singapore: Springer.

Jiang, H. & Hill, M. (2018). Teacher learning and classroom assessment in H. Jiang, & M. Hill (Eds.), *Teacher Learning within Classroom Assessment* (pp. 1-17). Singapore: Springer.

Jin, Y. (2010). The place of language testing and assessment in the professional preparation of foreign language teachers in China. *Language Testing, 27*(4), 555-584.

Johnson, J. C., & Weller, S. C. (2002). Elicitation techniques for interviewing. *Handbook of Interview Research: Context and Method*, 491-514.

Johnson, K. E. (2006). The sociocultural turn and its challenges for second language teacher education. *Tesol Quarterly*, *40*(1), 235-257.

Johnson, R., Becker, P., & Olive, F. (1999). Teaching the second-language testing course through test development by teachers-in-training. *Teacher Education Quarterly, 26*(3), 71-82.

Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher, 33*(7), 14-26.

Jupp, V. (2006). *The Sage dictionary of social research methods.* New York, NY: Sage.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kahn, E. A. (2000). A case study of assessment in a grade 10 English course. *The Journal of Educational Research*, *93*(5), 276-286.

Kaplan, D. (2009). *Structural equation modeling (2nd ed.).* San Francisco, CA: Sage.

Kealey, E. (2010). Assessment and evaluation in social work education: formative and summative approaches. *Journal of Teaching in Social Work, 30*(1), 64-74.

Kennedy, K. J. (2016). Exploring the influence of culture on assessment: The case of teachers' conceptions of assessment in Confucian Heritage Cultures. In G. T. Brown & L. R. Harris

(Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 404-419). New York, NY: Routledge.

Kennedy, K. J., Chan, J. K. S., & Fok, P. K. (2011). Holding policy-makers to account: exploring “soft” and “hard” policy and the implications for curriculum reform. *London Review of Education*, 9(1), 41-54.

Kleinsasser, R. (2005). Transforming a post-graduate level assessment course: a second language teacher educator's narrative. *Prospect*, *20*(3). 77-102.

Klem, L. (2000). Structural equation modeling. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding MORE multivariate statistics* (pp. 227-260). Washington, DC: American Psychological Association.

Kline, R. B. (2015). *Principles and practice of structural equation modelling*. New York, New York: Guilford Press.

Koh, K. H. (2011). Improving teachers’ assessment literacy through professional development. *Teaching Education*, 22(3), 255-276.

Kumaravadivelu, B. (2012). *Language teacher education for a global society: A modular model for knowing, analyzing, recognizing, doing, and seeing*. Routledge.

Lantolf, J. P. (Ed.). (2000). *Sociocultural theory and second language learning* (Vol. 78, No. 4). Oxford: Oxford university press.

Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing, 32*(2), 169-197.

Lam, R. (2016). Implementing assessment for learning in a Confucian context: The case of Hong Kong 2004–14. *The Sage handbook of curriculum, pedagogy and assessment, 2,* 756-771.

Lam, T. C., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement*, 317-322.

Lambert, D., & Lines, D. (2013). *Understanding assessment: Purposes, perceptions, practice*. UK: Routledge.

Lau, A. M. S. (2016). 'Formative good, summative bad?'–A review of the dichotomy in assessment literature. *Journal of Further and Higher Education, 40*(4), 509-525.

Leung, C. (2004). Developing formative teacher assessment: Knowledge, practice, and change. *Language Assessment Quarterly: An International Journal*, *1*(1), 19-41.

Li, M. (2004). Culture and classroom communication: A case study of Asian students in New Zealand language schools. *Asian EFL Journal, 6*(1), 275-303.

Li, W. S., & Hui, K. F. S. (2007). Conceptions of assessment of mainland China college lecturers: A technical paper analyzing the Chinese version of CoA-III. *The Asian Pacific Education Researcher, 16*(2), 185-198.

Liao, X. (2004). The need for communicative language teaching in China. *ELT Journal, 58*(3), 270-273.

Lieberman, A. (1995). Practices that support teacher development: Transforming conceptions of professional learning. *Innovating and Evaluating Science Education*, *2,* 67-78.

Light, G., Calkins, S., & Cox, R. (2009). *Learning and teaching in higher education: The reflective professional*. Thousand, Oaks: Sage.

Lin, L. C., Huang, P. H., & Weng, L. J. (2017). Selecting path models in SEM: A comparison of model selection criteria. *Structural Equation Modeling: A Multidisciplinary Journal, 24*(6), 855-869.

Liu, H. (2013). Reform of the college entrance examination: Ideology, principles, and policy recommendations. *Chinese Education & Society*, *46*(1), 10-22.

Liu, J., & Xu, Y. (2017). Assessment for learning in English language classrooms in China: Contexts, problems, and solutions. In H. Reinders, D. Nunan, & B. Zou (Eds.), *Innovation in Language Learning and Teaching, The Case of China* (pp. 17-37). London: Palgrave Macmillan.

Luo, R. (2003). Reflections on the implementation of classroom assessment in new curriculum reform. *Modern Education Science, 166* (2), 20-21.

Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing, 18*(4), 351-372.

Maibach, E., & Murphy, D. A. (1995). Self-efficacy in health promotion research and practice: conceptualization and measurement. *Health Education Research, 10*(1), 37-50.

Malone, M. E. (2008). Training in language assessment. In N.H. Hornberger, (Ed.), *Encyclopedia of Language and Education* pp (2362-2376). Boston, MA: Springer. Retrieved April, 2017, from https://doi.org/10.1007/978-0-387-30424-3.

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing*, *30*(3), 329-344.

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural equation models of latent interactions: evaluation of alternative estimation strategies and indicator construction. *Psychological Methods, 9*(3), 275.

May, K. A. (1991). Interview techniques in qualitative research: Concerns and challenges. In J. M. Moss (Ed.), *Qualitative Nursing Research: A contemporary Dialogue* (pp188-201), London, UK: Sage.

McGee, J., & Colby, S. (2014). Impact of an assessment course on teacher candidates' assessment literacy. *Action in Teacher Education*, *36*(5-6), 522-532.

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing*, *18*(4), 333-349.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20-32.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*(4), 203-213.

McMillan, J. H. (2013). Why we need research on classroom assessment. In J. H. McMillan *(Ed.), Sage Handbook of Research on Classroom Assessment (pp. 3-16)*. Thousand Oaks, CA: Sage.

Mei, B. (2016). *Towards an Understanding of Mainland Chinese Pre-Service EFL Teachers' Intention to Use CALL 2.0*. (Doctoral thesis), University of Auckland.

Mei, B., & Brown, G. T. (2018). Conducting online surveys in China. *Social Science Computer Review*, *36*(6), 721-734.

Mertens, D. (2015). *Research and evaluation in education and psychology: Integrating diversity with quantitative, qualitative, and mixed methods.* New York, NY: Sage Publications.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, *33*(1), 49-64.

Mertler, C. A. (2005). The role of classroom experience in preservice and in-service teachers' assessment literacy. *Midwest Educational Researcher, 18*(4), 25-33.

Mertler, C. A., & Campbell, C. (2005, April). *Measuring Teachers' Knowledge & Application of Classroom Assessment Concepts: Development of the "Assessment Literacy Inventory"*. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools*, *12*(2), 101-113.

Miao, Y. U. (2007). On the necessity of existence for grammar-translation method based on the present situation of English teaching in China. *Journal of Zhoukou Normal University, 4,* 35-46.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, *3*(1), 111-130.

Mitchell, R. (1992). *Testing for learning: How new approaches to evaluation can improve American schools.* New York, NY: Simon and Schuster.

Morgan, D. L. (1993). Qualitative content analysis: a guide to paths not taken. *Qualitative Health Research*, *3*(1), 112-121.

Morgan, D. L. (2014). Pragmatism as a paradigm for social research. *Qualitative Inquiry*, *20*(8), 1045-1053.

Moss, C. M. (2013). Research on classroom summative assessment. In J. H. MacMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 235-255). London, UK: Sage.

Munn, P., Drever, E., & Scottish Council for Research in Education. (1999). *Using questionnaires in small-scale research: A teacher's guide*, Scotland, UK: Scottish Council for Research in Education.

Nation, P. (2012). *What does every ESOL teacher need to know?* Closing plenary address at the 2012 CLESOL conference in Palmerston North. TESOLANZ, 20, 1-3.

Nelson, R., *&* Dawson, P. (2014). A contribution to the history of assessment: how a conversation simulator redeems Socratic method. *Assessment & Evaluation in Higher Education*. 39 (2): 195–204.

Neugebauer, R., & Wittes, J. (1994). Voluntary and involuntary capture-recapture samples--problems in the estimation of hidden and elusive populations. *American Journal of Public Health, 84*(7), 1068-1069.

Newton, J. (2009). A place for "intercultural" communicative language teaching (iCLT) in New Zealand ESOL classrooms. *TESOLANZ Journal*, *17*, 1-12.

New Zealand Ministry of Education. (2007). *The New Zealand curriculum.* Wellington, New Zealand: Learning Media.

O'Brien, J. G., Millis, B. J., & Cohen, M. W. (2009). *The course syllabus: A learning-centered approach* (Vol. 135). San Francisco: CA, John Wiley & Sons.

Odum, M. (2011, February). *Factor Scores, Structure and Communality Coefficients: A Primer.* Paper presented at the annual meeting of the Southwest Educational Research Association, San Antonio.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, *9*(6), 1-12.

O'Sullivan, B., & Stoynoff, S. (2012). *The Cambridge guide to second language assessment*. Cambridge, England: Cambridge University Press.

Palinkas, L. A., Aarons, G. A., Horwitz, S., Chamberlain, P., Hurlburt, M., & Landsverk, J. (2011). Mixed method designs in implementation research. *Administration and Policy in Mental Health and Mental Health Services Research, 38*(1), 44-53.

Palmer, D. H. (2006). Sources of self-efficacy in a science methods course for primary teacher education students. *Research in Science Education, 36*(4), 337-353.Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307-332.

Philips, D. (2000). Curriculum and assessment policy in New Zealand: Ten years of reforms. *Educational Review, 52*(2), 143-153.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher, 6*(1), 21-27.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice, 12*(4), 10-12.

Poole, A. (2016). Complex teaching realities and deep rooted cultural traditions: Barriers to the implementation and internalisation of formative assessment in China. *Cogent Education, 3*(1), 1156242.

Popham, W.J. (1999). "Why standardized tests don't measure educational quality". *Educational Leadership*. 56, 8-16.

Popham, W. J. (2004). Why Assessment Illiteracy Is Professional Suicide. *Educational Leadership, 62*(1), 82-96.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator, 46*(4), 265-273.

Popham, W. J. (2013). On serving two masters: formative and summative teacher evaluation. *Principal Leadership, 13*(7), 18-22.

Porter, A. C., & Gamoran, A. (2002). Progress and challenges for large-scale studies. In A.C. Porter & A. Gamoran (Eds.), *Methodological Advances in Cross-national Surveys of Educational Achievement*, (p 3). Washington, DC: National Academy Press.

Qu, W., & Zhang, C. (2013). The analysis of summative assessment and formative assessment and their roles in college English assessment system. *Journal of Language Teaching and Research*, *4*(2), 335.

Rappleye, J., & Komatsu, H. (2018). Stereotypes as Anglo-American exam ritual? Comparisons of students' exam anxiety in East Asia, America, Australia, and the United Kingdom. *Oxford Review of Education, 44*(6), 730-754.

Rea-Dickins, P. (2007). Classroom-based assessment: Possibilities and pitfalls. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 505-520). Boston, MA: Springer.

Richards, J. C., & Renandya, W. A. (Eds.). (2002). *Methodology in language teaching: An anthology of current practice*. Cambridge, UK: Cambridge university press.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge, UK: Cambridge University Press.

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), 242-258.

Rowley, J. and Slack, F. (2004), Conducting a literature review, *Management Research News, 27* (6), 31-39.

Rushton, A. (2005). Formative assessment: a key to deep learning?. *Medical Teacher*, *27*(6), 509-513.

Sadler, D. R. (1998). Formative assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 77-84.

Sapp, M. (2013). *Test anxiety: Applied research, assessment, and treatment interventions.* Plymouth, UK: University Press of America.

Savignon, S. J. (1991). Communicative language teaching: State of the art. *TESOL Quarterly*, *25*(2), 261-278.

Savignon, S. J. (Ed.). (2002). *Interpreting communicative language teaching*. New Haven, CT: Yale University Press.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research*, *99*(6), 323-338.

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), *Perspectives of Curriculum Evaluation* (Vol. 1, pp. 39-55). Chicago: Rand McNally.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Shepard, L.A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Westport, CT: Praeger.

Shohamy, E. (2007). Language tests as language policy tools. *Assessment in Education*, *14*(1), 117-130.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' assessment literacy. *Journal of Science Teacher Education, 22*(4), 371-391.

Singh, G., & Richards, J. C. (2006). Teaching and learning in the language teacher education course room: A critical sociocultural perspective. *RELC Journal, 37*(2), 149-175.

Singh, R. (2009). Does my structural model represent the real phenomenon?: A review of the appropriate use of Structural Equation Modelling (SEM) model fit indices. *The Marketing Review*, *9*(3), 199-212.

Slattery, J. M., & Carlson, J. F. (2005). Preparing an effective syllabus: Current best practices. *College Teaching*, *53*(4), 159-164.

Smith, C. P. (2000). Content analysis and narrative analysis. In H. T. Reis & C. M. Judd (Eds.), *Handbook of Research Methods in Social and Personality Psychology* (pp. 313-335). Cambridge, UK: Cambridge University Press.

Smith, H. J., Chen, J., & Liu, X. (2008). Language and rigour in qualitative research: problems and principles in analyzing data collected in Mandarin. *BMC Medical Research Methodology*, *8*(1), 44.

Smith, L. F., Hill, M. F., Cowie, B., & Gilmore, A. (2014). Preparing teachers to use the enabling power of assessment. In *Designing Dssessment for Quality Learning* (pp. 303-323). Springer, Dordrecht.

Spolsky, B. (1995). *Measured words: The development of objective language testing.* Oxford: Oxford University Press.

Stajkovic, A. D., & Luthans, F. (1998). Self-efficacy and work-related performance: A meta-analysis. *Psychological Bulletin*, 124(2), 240.

Stanny, C., Gonzalez, M., & McGowan, B. (2015). Assessing the culture of teaching and learning through a syllabus review. *Assessment & Evaluation in Higher Education*, *40*(7), 898-913.

Steinmetz, H., Schmidt, P., Tina-Booh, A., Wieczorek, S., & Schwartz, S. H. (2009). Testing measurement invariance using multigroup CFA: Differences between educational groups in human values measurement. *Quality & Quantity*, 43(4), 599.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*(7), 534-39.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-243.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan, 83*(10), 758-765.

Stiggins, R. J., & Conklin, N. F. (1992). *In teacher's hands: Investigating the practices of classroom assessment*. Albany, NY: Suny Press.

Sue, V. M., & Ritter, L. A. (2012). *Conducting online surveys*. Thousand Oaks, CA: Sage.

Suen, L. J. W., Huang, H. M., & Lee, H. H. (2014). A comparison of convenience sampling and purposive sampling. *The Nursing Magazine*, *61*(3), 105.

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies, 53*(4), 466-478.

Taylor, L. (2005). Washback and impact. *ELT Journal*, *59*(2), 154-155.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29,* 21-36.

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35-39.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioural sciences*. NY: Sage.

Tesch, R. (2013). *Qualitative research: Analysis types and software*. New York, NY: Routledge.

Tongco, M. D. C. (2007). Purposive sampling as a tool for informant selection. *Ethnobotany Research and Applications*, *5*, 147-158.

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: Research, challenges and future prospects. *Language Testing and Assessment*, *6*(1), 41-63.

Ullman, J. B. (2006). Structural equation modelling: Reviewing the basics and moving forward. *Journal of Personality Assessment*, *87*(1), 35-50.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70.

Volante, L., & Fazio, X. (2007). Exploring Teacher Candidates' Assessment Literacy: Implications for Teacher Education Reform and Professional Development. *Canadian Journal of Education*, *30*(3), 749-770.

Wang, W., & Lam, A. S. (2009). The English language curriculum for senior secondary school in China: Its evolution from 1949. *RELC Journal*, *40*(1), 65-82.

Wang, X. (2009). *The Process of English Language Teaching and Learning experienced by Teachers and Students in Two Contexts: China and New Zealand* (Doctoral dissertation, The University of Waikato).

Wang, X. (2017). A Chinese EFL Teacher's Classroom Assessment Practices. *Language Assessment Quarterly, 14*(4), 312-327.

Watkins, D., & Gioia, D. (2015). *Mixed methods research: Pocket guides to social work research guide.* Oxford: Oxford University Press.

Wiliam, D., & Black, P. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment. *British Educational Research Journal, 22*(5), 537-548.

Wiliam, D. (2011). What is assessment for learning?. *Studies in Educational Evaluation*, *37*(1), 3-14.

Wharton, S. (1998). Teaching language testing on a pre-service TEFL course. *ELT Journal, 52*(2), 127-132.

Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with TIMSS data. *Practical Assessment, Research & Evaluation*, *12*(3), 1-26.

Wu, Z. (2011). *Preservice teacher education: case study of BA/TEFL curriculum development*. Shanghai Foreign Language Education Press, Shanghai: China.

Xu, S., & Connelly, F. M. (2009). Narrative inquiry for teacher education and development: Focus on English as a foreign language in China. *Teaching and Teacher Education*, *25*(2), 219-227.

Xu, Y., & Brown, G. T. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149-162.

Xu, Y., & Brown, G. T. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment*, *6*(1), 133-158.

Yan, C., & He, C. (2010). Transforming the existing model of teaching practicum: A study of Chinese EFL student teachers' perceptions. *Journal of Education for Teaching*, *36*(1), 57-73.

Yin, R. K. (2010). *Qualitative research from start to finish*. New York, NY: The Guilford Press.

Yu, L., & Suen, H. K. (2005). Historical and contemporary exam-driven education fever in China. *KEDI Journal of Educational Policy*, 2(1), 17-33.

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*(1), 115-148.

Zhan, S. (2008). Changes to a Chinese pre- service language teacher education program: Analysis, results and implications. *Asia- Pacific Journal of Teacher Education*, *36*(1), 53-70.

Zhao, D., Yan, B., Tang, L., & Zhou, Y. (2018). Teachers' Understanding, Implementing and Reflecting upon Classroom Assessment in Primary and Junior High Schools of China. In *Teacher Learning with Classroom Assessment* (pp. 77-98). Singapore: Springer.

Zhou, G & Zhou, X. (2019). *Education policy and reform in China*. Singapore: Palgrave Macmillan.

Appendices

Appendix A

Phase-I Data Collection PIS and CF

PARTICIPANT INFORMATION SHEET

COURSE INSTRUCTOR

Dear participant,

My name is Wenjing (Jenny) Yao, a PhD candidate in the School of Learning, Development and Professional Practice, Faculty of Education, the University of Auckland. I sincerely invite you to participate in this research project as part of my doctoral thesis. The following information highlights the basic information about this research so that you can read and make an informed decision as to whether you would like to participate or not. The entire research project will be guided and supervised by the researcher's supervisors, Associate Professors Mary Hill and Gavin Brown.

Project title
Language assessment versus language testing- a comparative study of the language assessment courses in New Zealand and China

Project description and procedures
This research project will investigate the language assessment courses provided by colleges and universities in New Zealand and China. As the main researcher of this study, I firmly believe that your insight will be of great help to understand and address the challenges facing language assessment and testing courses in New Zealand and China. Thus I sincerely invite you to participate in this project. Altogether a maximum 60-minute semi-structured interview and one period of classroom observation will be conducted. The interview questions are mainly about your conceptions of assessment and the observation is about your assessment practices. The interview will be audio-recorded and the classroom observation video-recorded (with your permission) but you have the right to ask the researcher to stop recording anytime during the interview or the observation. If you do not wish to be recorded, the researcher will take notes during the interview or the observation. During the classroom observation, the researcher will focus on, and only record, the lecturer. Should students be asked to answer questions, deliver presentations or do other class activities, the researcher would turn off the recording so as to avoid any incidental recording of the students' voices or behaviour. There is no potential harm in the interview or the observation and all the information will be kept confidential, but if you want to withdraw from the process you can do so at any time. If you agree to participate in my research, a thank-you gift will be offered as a token of my sincere gratitude for your support and participation.

Data storage, retention, destruction and future use
Data will be stored and protected either in the researchers' personal password protected computer or in her locked office cabinet. Data will be stored for six years after completion of the research, and deleted after this time, unless the researcher continues with research in this field. Data collected will only be used for academic and educational purposes, such as the researcher's PhD thesis at the University of Auckland, academic publications, conference presentations, teaching, and other forms of academic research dissemination.

Participants' rights
It's completely up to you to decide whether you will take part in this research. You can ask to receive the electronic copies of your interview and observation recordings and transcripts via email or cloud-delivery. You are able to change, delete and add further information on drafts of the transcripts.

Anonymity and Confidentiality

Your identity and information will remain confidential during this research and no personal features will be identified. You can choose a pseudonym in the attached Consent Form; otherwise the researcher will make up one for you.

Thank you for considering participation in this study.

Contact details

All researchers are at the Faculty of Education, The University of Auckland, 74 Epsom Avenue, Epsom, Auckland 1023, New Zealand.

Researcher	Main supervisor	Co-supervisor
Wenjing (Jenny) Yao	Associate Professor Gavin Brown	Associate Professor Mary Hill
+64 021 2101083 (NZ) +86 13611885162(PRC)	+64 9 373 7599 ext. 48602	+64 9 373 7599 ext. 48630
w.yao@auckland.ac.nz	gt.brown@auckland.ac.nz	hill.m@auckland.ac.nz

For any ethical concerns please contact: The Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Private Bag 92019, Auckland, 1142. Telephone 09 373-7599 ext. 87830/83761. Email: humanethics@auckland.ac.nz.

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 25/03/2015 FROM 03/25/2015 TO 03/25/2018. REFERENCE NUMBER 014023.

PARTICIPANT CONSENT FORM

COURSE INSTRUCTOR

THIS FORM WILL BE HELD FOR A MINIMUM PERIOD OF 6 YEARS

Project title: Language assessment versus language testing-a comparative study of language assessment courses in New Zealand and China

Researcher: Wenjing (Jenny) Yao

I have read the Participant Information Sheet and understood the nature of the research and why I have been invited to participate in this research project. I have had the opportunity to ask questions and have them answered.

1. I am aware that the researcher will conduct a 60-minute face-to-face interview and one period of classroom observation.
2. I understand that I can refuse to answer interview questions, stop the interview, or choose to withdraw from the research at any time.
3. I understand that I may be audio-recorded during the interview and video-recorded during the observation. If I do not wish to be recorded, the researcher will take notes during the interview or the observation.

   Do you agree to the interview being audio recorded? YES/ NO

   Do you agree to being video-recorded during the lesson? YES/NO

   I understand that the recordings will be transcribed by the researcher.
3. I wish to receive the electronic copies of my recordings and transcripts via email or cloud-delivery. YES/NO
4. I understand that I am able to change, delete and add further information on drafts of the transcripts.
5. I understand that the data will be used for the researcher's PhD thesis and may be used for academic publications, conference presentations, teaching and other forms of academic research dissemination.
6. I understand that the data can only be accessed by the researcher and her supervisors.
7. I understand that the data will be stored separately from this form and securely for a period of six years and then deleted, unless the researcher continues with this line of research.
8. I wish to receive the final copy of research findings. YES?NO (If YES, please indicate whether you want to receive this by email or cloud delivery and provide an address for this ____________
9. I understand that I can choose a pseudonym for the research or will be given one by the researcher. No real names will be used and no personal information will be revealed.

My email: ___________________________ Pseudonym: _____________________________

Name ____________________________ Signature: ___________________________ Date: ________________

For any ethical concerns please contact: The Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Private Bag 92019, Auckland, 1142. Telephone 09 373-7599 ext. 87830/83761. Email: humanethics@auckland.ac.nz

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 25/03/2015 FROM 03/25/2015 TO 03/25/2018. REFERENCE NUMBER 014023.

PARTICIPANT INFORMATION SHEET

DEAN

Dear Dean,

This is Wenjing (Jenny) Yao, a PhD candidate in the School of Learning, Development and Professional Practice, Faculty of Education, the University of Auckland. I sincerely ask for your permission for me to approach and invite staff in the Faculty of Arts to participate in this research project as part of my doctoral thesis. The following information highlights the basic information about this research so that you can read and make an informed decision. The entire research project will be guided and supervised by the researcher's supervisors, Associate Professors Mary Hill and Gavin Brown.

Project title
Language assessment versus language testing- a comparative study of the language assessment courses in New Zealand and China

Project description and procedures
This research project will investigate the language assessment courses provided by universities in New Zealand and China with the focus to discover and analyse the challenges facing language assessment courses in NZ and PRC and propose feasible solutions. Altogether a maximum 60-minute semi-structured interviews with the course lecturer/s and one period of classroom observation during a language assessment course will be carried out in the participating university. The interview will be audio-recorded and the classroom observation of the lecturer video-recorded. During the classroom observation, the researcher will focus on and only record, the lecturer/s to avoid any incidental recording of the students' voices or behaviour. There is no potential harm in the interview or the observation and all the information will be kept confidential, and the participant/s can withdraw from the research at any time.

Participants of this research will be informed of this study with a participation information sheet and ask for their consent to participate in the consent form and they are free to choose to participate or decline. Information about the participants will remain confidential during and after this research and no personal features will be identified.

We would really appreciate if you approve our access to the site of the Faculty of Arts and approach your staff for this research.

Contact details
All researchers are at the Faculty of Education, The University of Auckland, 74 Epsom Avenue, Epsom, Auckland 1023, New Zealand.

Researcher	Main supervisor	Co-supervisor
Wenjing (Jenny) Yao	Associate Professor Gavin Brown	Associate Professor Mary Hill
+64 021 2101083 (NZ) +86 13611885162(PRC)	+64 9 373 7599 ext. 48602	+64 9 373 7599 ext. 48630
w.yao@auckland.ac.nz	gt.brown@auckland.ac.nz	hill.m@auckland.ac.nz

For any ethical concerns please contact: The Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Private Bag 92019, Auckland, 1142. Telephone 09 373-7599 ext. 87830/83761. Email: humanethics@auckland.ac.nz.

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 25/03/2015 FROM 03/25/2015 TO 03/25/2018. REFERENCE NUMBER 014023.

DEAN CONSENT FORM

THIS FORM WILL BE HELD FOR A MINIMUM PERIOD OF 6 YEARS

Project title: Language assessment versus language testing-a comparative study of language assessment courses in New Zealand and China

Researcher: Wenjing (Jenny) Yao, PhD candidate supervised by Associate Professors Mary Hill and Gavin Brown in the Faculty of Education, at the University of Auckland.

1. I have been provided an explanation of this research and understood its nature.
2. I have had an opportunity to ask questions and have them answered.
3. I give consent for the researcher to invite participants at this site and to conduct research as described in the participant information sheet.
4. I confirm that whether the Faculty member/s choose to participate or not, this will have no influence on their employment status.

Name ___________________ Signature: ____________________ Date: _____________

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 25/03/2015 FROM 03/25/2015 TO 03/25/2018. REFERENCE NUMBER 014023.

Appendix B

Interview Questions

1. Would you briefly describe your language assessment/testing course? (Prompt: Say, the course objectives, content, requirement, assessment and textbooks or teaching materials in particular).

   请简要描述下您的语言测评课程（如教学目标、内容、要求、评估方式和选用教材等）。

2. What qualities do you think an assessment literate teacher should possess? (Prompt: why are these essential?)

   您认为，一个有测评素养的教师应该具备什么样的素质？

3. What do you think is the relationship between language testing and other forms of language assessment? (Prompt: do you think testing should be the major form of assessment?)

   您认为语言测试和其他种类的评估有什么样的关系？测试应该是主要评估手段吗？

4. What kind of assessment activities and tools do you mainly apply in your teaching? 在教学中您使用哪些测评活动、方法和工具？

5. How do you build students’ assessment ability in the teaching?

   您在教学中怎样发展学生们的测评能力？

6. How do you help students practice what they have learned in the course? Either in the classroom or outside the classroom.

   您在教学中怎样帮助学生实践所学到的知识？（课堂外和课堂上）

7. In your opinion, what are the challenges facing language teacher preparation with regard to their assessment ability? (Prompt: What potential is there for change?)

   在培养学生的测评能力中，您遇到的挑战有哪些？

8. Are there any aspects in your teaching that you would like to change? Why and how?

   在教学中您有想改动的方面吗？为什么？怎样改？

9. What role, if any, do you think policy (national level or university faculty level) plays in the construction of your course?

   您认为国家和学校现行的教学和评估方面的政策对您的课程有影响吗？

10. Do you think the current socio-cultural or educational environment have some impact on your course?

    您认为现有的社会文化和教育环境对您的课程有哪些影响？

Appendix C

Phase-II Data Collection PIS and CF

PARTICIPANT INFORMATION SHEET

TEACHERS AND TUTORS

Project title: Investigating Chinese pre-service EFL teachers' assessment literacy---Do language assessment courses make a difference?
Research team: Wenjing (Jenny) Yao, Professor Gavin Brown and Associate Professor Mary Hill.

You are invited to assist in a survey of how pre-service EFL (English as foreign language) teachers in China develop and establish assessment literacy with or without the language assessment courses. This research will help us understand how pre-service teachers develop assessment capability and potentially point to new ways of delivering language assessment education in China.

We ask for your help in alerting students, i.e. the pre-service teachers to the survey by sharing either by email, WeChat, QQ, or in print the advertisement we have prepared. The 20-minute online survey asks students about their opinions and experiences around language assessment. No IP addresses or other identifying information will be gathered. This means we will not be able to tell you if your students participated and what answers they gave. Information about the students will be kept anonymous and confidential. Participation is completely voluntary. As the survey is anonymous, submission of the survey counts as students' consent to participate and they will not be able to withdraw their data once they submit their questionnaire. By completing the survey, students can enter a draw to win prizes of portable chargers or U disks worth about RMB50. The potential number of student participants is no less than 1000.

All the data from this survey will be stored and protected on password protected computers and secure servers at the University of Auckland for at least six years. Data collected will only be used for academic and educational purposes, such as the student researcher's PhD thesis at the University of Auckland, academic publications, conference presentations, teaching, and other forms of academic research dissemination.

Should you have any concerns during the process, you can contact the following people for clarification and you are welcome to ask for publications based on this study by contacting the student researcher.

Student researcher	**Main supervisor**	**Co-supervisor**	**Head of School**
Wenjing (Jenny) Yao	Professor Gavin Brown	Assoc. Prof Mary Hill	Assoc. Prof Richard Hamilton
Wechat: 8178606	+64 9 373 7599 ext. 48602	+64 9 373 7599 ext. 48630	09 373- 7999 ext. 85619
w.yao@auckland.ac.nz	gt.brown@auckland.ac.nz	hill.m@auckland.ac.nz	r.hamilton@auckland.ac.nz

For any ethical concerns please contact: The Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Private Bag 92019, Auckland, 1142. Telephone 09 373-7599 ext. 83711. Email: humanethics@auckland.ac.nz.

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 15/07/2016 FROM 15/07/2019 TO 03/25/2018. REFERENCE NUMBER 019469.

PARTICIPANT INFORMATION SHEET

PRE-SERVICE TEACHERS

Project title: Investigating Chinese pre-service EFL teachers' assessment literacy---Do language assessment courses make a difference?
Research team: Wenjing (Jenny) Yao, Professor Gavin Brown and Associate Professor Mary Hill.

You are invited to participate in a survey research of how pre-service EFL (English as foreign language) teachers in China develop and establish their assessment knowledge and skills with or without the help of language assessment courses. Your participation will possibly help improve teacher education in China.

The survey is completely anonymous. No IP addresses or other identifying information will be gathered. Your teachers will not be told if you participated or how you answered. All responses are fully protected using industry standard methods. Participation in this survey is completely voluntary. To participate, follow the link or scan the QR code provided in the research advertisement to connect to the survey. We estimate this will take no more than 20 minutes. Please note that your submission counts as consent to participate and data cannot be withdrawn once the survey has been completed. The number of potential participants is no less than 1000.

At the end of the survey, you can choose to enter a lottery to win a prize worth RMB50 (either a portable charger or a U disks). In order to enter, click on the link at the end of the survey to go to a new page where you can leave your name and email address. This contact information will be stored separately to your answers, so we won't know how you answered. You can request a copy of the report of this survey and it will be sent to you once it has been finalised.

We hope that participating in this survey will help you gain a more thoughtful understanding of language assessment. The research procedures are being carried out by the student investigator Wenjing (Jenny) Yao who is supervised by Professor Gavin Brown and Associate Professor Mary Hill.

All the data from this study will be stored on password-protected computers and protected by secure servers at the University of Auckland for six years. Data collected will only be used for academic and educational purposes, such as the student researcher's PhD thesis, academic publications, conference presentations, teaching, and other forms of academic research dissemination. Should you have any concerns about the process, you can contact the following people for clarification who can be reached at the address above.

Student researcher	**Main supervisor**	**Co-supervisor**	**Head of School**
Wenjing (Jenny) Yao	Professor Gavin Brown	Assoc. Prof Mary Hill	Assoc. Prof Richard Hamilton
Wechat: 8178606	+64 9 373 7599 ext. 48602	+64 9 373 7599 ext. 48630	09 373- 7999 ext. 85619
w.yao@auckland.ac.nz	gt.brown@auckland.ac.nz	hill.m@auckland.ac.nz	r.hamilton@auckland.ac.nz

For any ethical concerns please contact: The Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Private Bag 92019, Auckland, 1142. Telephone 09 373-7599 ext. 83711. Email: humanethics@auckland.ac.nz.

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON 15/07/2016 FROM 15/07/2019 TO 03/25/2018. REFERENCE NUMBER 019469.

Appendix D

Conceptions, Self-efficacy and Practice of Assessment Questionnaire-English version

Q1 PARTICIPANT INFORMATION SHEET

Project title: Investigating Chinese pre-service EFL teachers' assessment literacy---Do language assessment courses make a difference?

Research team: Wenjing (Jenny) Yao, Professor Gavin Brown and Associate Professor Mary Hill.

You are invited to participate in a survey research of how pre-service EFL (English as foreign language) teachers in China develop and establish their assessment knowledge and skills with or without the help of language assessment courses. Your participation will possibly help improve teacher education in China.

The survey is completely anonymous. No IP addresses or other identifying information will be gathered. Your teachers will not be told if you participated or how you answered. All responses in the Qualtrics software are fully protected using industry standard methods. Participation in this survey is completely voluntary. To participate, follow the link or scan the QR code provided in the research advertisement to connect to the survey. We estimate this will take no more than 20 minutes. Please note that your submission counts as consent to participate and data cannot be withdrawn once the survey has been completed. The number of potential participants is no less than 1000.

At the end of the survey, you can choose to enter a lottery to win a prize worth RMB50 (either a portable charger or a U disk). In order to enter, click on the link at the end of the survey to go to a new page where you can leave your name and email address. This contact information will be stored separately to your answers, so we won't know how you answered. You can request a copy of the report of this survey and it will be sent to you once it has been finalised.

We hope that participating in this survey will help you gain a more thoughtful understanding of language assessment. The research procedures are being carried out by the student investigator Wenjing (Jenny) Yao who is supervised by Professor Gavin Brown and Associate Professor Mary Hill.

All the data from this study will be stored on password-protected computers and protected by secure servers at the University of Auckland for six years. Data collected will only be used for academic and educational purposes, such as the student researcher's PhD thesis, academic publications, conference presentations, teaching, and other forms of academic research dissemination. Should you have any concerns about the process, you can contact the following people for clarification who can be reached at the address above.

Should you have any concerns about the process, you can contact the following people for clarification who can be reached at the address above.

Student researcher	**Main supervisor**	**Co-supervisor**	**Head of School**
Wenjing (Jenny) Yao	Professor Gavin Brown	Assoc. Prof Mary Hill	Assoc. Prof Richard Hamilt
Wechat: 8178606	+64 9 373 7599 ext. 48602	+64 9 373 7599 ext. 48630	09 373- 7999 ext. 85619
w.yao@auckland.ac.nz	gt.brown@auckland.ac.nz	hill.m@auckland.ac.nz	r.hamilton@auckland.ac.nz

All researchers are at the Faculty of Education, The University of Auckland, 74 Epsom Avenue, Epsom, Auckland 1023, New Zealand. For any ethical concerns please contact: The Chair, The University of Auckland Human Participants Ethics Committee, The University of Auckland, Private Bag 92019, Auckland, 1142. Telephone 09 373-7599 ext. 87830/83761. Email: humanethics@auckland.ac.nz

APPROVED BY THE UNIVERSITY OF AUCKLAND HUMAN PARTICIPANTS ETHICS COMMITTEE ON ………… for (3) years, Reference Number HPEC 2017/019496

Q2 Are you a male or a female?

❍ Female (1)
❍ Male (2)
❍ Choose not to answer (3)

Q3 Which year of college are you in?

❍ year 3 (1)
❍ year 4 (2)

Q4 At which university are you currently enrolled in? (The universities are listed in alphabetical order)

❍ Anhui Normal University (1)
❍ Beijing Normal University (2)
❍ Central China Normal University (3)
❍ East China Normal University (4)
❍ Guangzhou Foreign Studies University (5)
❍ Hangzhou Normal University (6)
❍ Henan Normal University (7)
❍ Langfang Teacher College (8)
❍ Nanjing Normal University (9)
❍ Northeast Normal University (10)
❍ Shanghai Normal University (11)
❍ Shanghai Foreign Studies University (12)
❍ West China Normal University (13)
❍ Zhejiang Normal University (14)
❍ Other (15) ____________________

Q5 Have you taken any language assessment course/s?

❍ No (1)
❍ Yes (2)

Q6 What was the name of the course you took?

❍ Language testing (1)
❍ Language testing and assessment (2)
❍ Other (3) ____________________

Now, you will be asked to answer items about assessment and how you would like to carry it out in your own teaching. In all items the word ‘assessment’ refers to the activities that teachers use to gather information about students’ learning progress, attitudes, problems, etc. to improve their teaching and students' learning.

Q7 In this section, please indicate how strongly you agree or disagree with the following statements.	Strongly Disagree (1)	Mostly Disagree (2)	Slightly Agree (3)	Moderately Agree (4)	Mostly Agree (5)	Strongly Agree (6)
Assessment helps students succeed in real-world experiences. (1)	❍	❍	❍	❍	❍	❍
Assessment fosters students' character. (2)	❍	❍	❍	❍	❍	❍
Assessment is used to provoke students to be interested in learning. (3)	❍	❍	❍	❍	❍	❍
Assessment stimulates students to think. (4)	❍	❍	❍	❍	❍	❍
Assessment cultivates students' positive attitudes towards life. (5)	❍	❍	❍	❍	❍	❍
Assessment helps students improve their learning. (6)	❍	❍	❍	❍	❍	❍
Assessment determines if students meet qualification standards. (7)	❍	❍	❍	❍	❍	❍
Assessment information modifies ongoing teaching of students. (8)	❍	❍	❍	❍	❍	❍
Assessment interferes with teaching. (9)	❍	❍	❍	❍	❍	❍
Assessment is an imprecise process. (10)	❍	❍	❍	❍	❍	❍
Assessment has little	❍	❍	❍	❍	❍	❍

impact on teaching. (11)						
Assessment forces teachers to teach in a way against their beliefs. (12)	❍	❍	❍	❍	❍	❍
Assessment indicates how good a teacher is. (13)	❍	❍	❍	❍	❍	❍
Assessment is an accurate indicator of a school's quality. (14)	❍	❍	❍	❍	❍	❍
Assessment measures the worth or quality of schools. (15)	❍	❍	❍	❍	❍	❍
Assessment is used by school leaders to police what teachers do. (16)	❍	❍	❍	❍	❍	❍
Teachers should take into account error and measure imprecision in all assessment (17)	❍	❍	❍	❍	❍	❍
Assessment results are trustworthy. (18)	❍	❍	❍	❍	❍	❍
Assessment results can be depended on. (19)	❍	❍	❍	❍	❍	❍
Assessment results are sufficiently accurate. (20)	❍	❍	❍	❍	❍	❍
Assessment results are filed and ignored. (21)	❍	❍	❍	❍	❍	❍
Assessment helps students gain good scores in	❍	❍	❍	❍	❍	❍

examinations. (22)						
Assessment familiarizes students with examination formats. (23)	❍	❍	❍	❍	❍	❍
Assessment teaches examination-taking techniques. (24)	❍	❍	❍	❍	❍	❍
Assessment prepares students for examinations. (25)	❍	❍	❍	❍	❍	❍
Assessment helps students avoid failure on examinations. (26)	❍	❍	❍	❍	❍	❍
Assessment is assigning a grade or level to students work. (27)	❍	❍	❍	❍	❍	❍
Assessment sets the schedule or timetable for classes. (28)	❍	❍	❍	❍	❍	❍
Assessment selects students for future education or employment opportunities. (29)	❍	❍	❍	❍	❍	❍
Assessment results should be treated cautiously because of measurement error. (30)	❍	❍	❍	❍	❍	❍
Assessment results contribute to teachers' appraisals. (31)	❍	❍	❍	❍	❍	❍

Q8 How sure are you that you can carry out the following practices?	Very unsure (1)	Mostly unsure (2)	Slightly sure (3)	Moderately sure (4)	Mostly sure (5)	Very Sure (6)
I'm able to select the right tests for my students. (1)	❍	❍	❍	❍	❍	❍
I'm able to write the right tests for my students. (2)	❍	❍	❍	❍	❍	❍
I'm able to carry out formative assessment activities in the classroom/in my teaching. (3)	❍	❍	❍	❍	❍	❍
I'm able to use a variety of assessment methods in teaching. (4)	❍	❍	❍	❍	❍	❍
I'm able to align my assessment activities with my teaching goals. (5)	❍	❍	❍	❍	❍	❍
I'm able to apply what I've learned in practical teaching. (6)	❍	❍	❍	❍	❍	❍
I'm able to interpret assessment results appropriately. (7)	❍	❍	❍	❍	❍	❍
I'm able to provide students with assessment results in an appropriate way. (8)	❍	❍	❍	❍	❍	❍

Q9 How likely are you to carry out these classroom assessment practices in your future teaching?	Very unlikely (1)	Moderately unlikely (2)	Slightly likely (3)	Moderately likely (4)	Likely (5)	Very likely (6)
I point out areas for improvement to students when I give them feedback (1)	❍	❍	❍	❍	❍	❍
I share learning goals and success criteria with students to help their learning. (2)	❍	❍	❍	❍	❍	❍
I encourage and value students’ own evaluation of their work. (3)	❍	❍	❍	❍	❍	❍
I use portfolio assessment to record students’ learning process. (4)	❍	❍	❍	❍	❍	❍
I ask questions in class to check students’ understanding. (5)	❍	❍	❍	❍	❍	❍
I observe students’ classroom performance and base my teaching on the strengths and weakness I see. (6)	❍	❍	❍	❍	❍	❍
I regularly ask students to give feedback on each other’s work (7)	❍	❍	❍	❍	❍	❍
I interact with the students to learn what I need to teach next. (8)	❍	❍	❍	❍	❍	❍

I check that students understand and use my feedback. (9)	❍	❍	❍	❍	❍	❍
I monitor student peer assessments to make sure they give each other useful comments. (10)	❍	❍	❍	❍	❍	❍
I emphasize constructive comments over giving scores or grades. (11)	❍	❍	❍	❍	❍	❍
I assess according to national or provincial standards or norms. (12)	❍	❍	❍	❍	❍	❍
I assess according to *zhongkao/ gaokao* rules. (13)	❍	❍	❍	❍	❍	❍
I try my best to help students get high test scores. (14)	❍	❍	❍	❍	❍	❍
I regularly practice tests with students to help them prepare for final exams. (15)	❍	❍	❍	❍	❍	❍
I use classroom assessment to help students prepare for tests. (16)	❍	❍	❍	❍	❍	❍
The priority of my assessment activities is to help students get good results in their examinations. (17)	❍	❍	❍	❍	❍	❍
I value students’ test results more	❍	❍	❍	❍	❍	❍

than their classroom performance. (18)						
I use tests designed primarily by myself. (19)	○	○	○	○	○	○
I prefer getting tests from an external source rather than writing them myself. (20)	○	○	○	○	○	○
I look at test results to reflect on and improve my teaching. (21)	○	○	○	○	○	○
I review test results to diagnose students' learning problems. (22)	○	○	○	○	○	○
Tests are an important way to motivate students to learn. (23)	○	○	○	○	○	○
I depend on test results to know what students have learned. (24)	○	○	○	○	○	○
My tests are aligned with teaching goals. (25)	○	○	○	○	○	○

Q10 Would you like to take part in a draw to win prizes? (a portable charger or a U disk worth RMB 50 for every 50 students)

○ No (2)
○ Yes (1)

Note: If participants click Yes, they will be directed to another survey as below:

Q1 What is your name?

Q3 What is your email address?

Q2 Choose the prize you want to win

○ Portable charger (1)
○ 16G U disk (2)

Conceptions, Self-efficacy and Practice of Assessment Questionnaire-Chinese version

英语专业职前教师评估素养问卷（观念，信念及实践）

你好,感谢你打开这个链接。我是一名英语教师，也是一名在读博士生。我的研究想了解英语教师（职前或在职）的评估观和评估实践。问卷为匿名，答卷约需8分钟。完成问卷后你会对教学中的评估有更深刻的理解，也有机会参加抽奖获得奖品。你的回答对我的研究很重要，请根据实际情况认真回答。

如有任何疑问，请联系：

姚雯静（博士在读） Wechat: 8178606, email: w.yao@auckland.ac.nz

Professor Gavin Brown（主导师） gt.brown@auckland.ac.nz

Assoc. Prof. Mary Hill （副导师）hill.m@auckland.ac.nz

非常感谢!

你是英语师范类或英语教育专业学生吗？ [单选题] *

○是 (请跳至第2题)

○不是 (请跳至第3题)

你是大三大四学生吗？ [单选题] *

○是 (请跳至第4题)

○不是 (请跳至第3题)

你的性别： [单选题] *

○ ○

男 女

你就读于哪所学校？ [填空题] *

____

你是几年级学生？ [单选题] *

○大三

○大四

你上过“语言测试(language testing)”或“语言测评(language testing and assessment)”课吗？ [单选题] *

○有

○没有

你上过的课程名称是？ [单选题] *

○语言测试(Language testing)

○语言测试和评估(Language testing and assessment)

○其他 _________________

问卷中的“评估”指教学中教师通过各种方式收集和了解学生学习情况和学习态度等方面的教学活动。请结合你的情况回答下列问题。请阅读下列句子，选择你认同的程度。从左到右分别为，两个“不同意”选项，四个“同意”选项。[矩阵量表题] *

	1 非常不同意	2 不同意	3 有点同意	4 同意	5 很同意	6 非常同意
评估有助学生改善学习。	○	○	○	○	○	○
评估帮助学生在实际生活中获得成功。	○	○	○	○	○	○
评估培养学生性格。	○	○	○	○	○	○
评估用来激发学生学习兴趣。	○	○	○	○	○	○

评估激励学生思考。	○	○	○	○	○	○
评估培养学生树立积极的人生观。	○	○	○	○	○	○
评估评定学生是否达到标准。	○	○	○	○	○	○
评估信息有助于教师改进教学。	○	○	○	○	○	○
评估干扰教学。	○	○	○	○	○	○
评估是一个不精确的过程。	○	○	○	○	○	○
评估对教学影响微不足道。	○	○	○	○	○	○
评估迫使教师用违背自己教学理念的方法教学。	○	○	○	○	○	○
评估反映出教师是否称职。	○	○	○	○	○	○
评估结果是检验学校质量的准确指标。	○	○	○	○	○	○
评估衡量学校的价值和质量。	○	○	○	○	○	○
学校领导利用评估来管	○	○	○	○	○	○

制教师做所的事。						
教师在评估活动中应考虑到评估的误差和不精确性。	○	○	○	○	○	○
评估帮助学生熟悉考试题目。	○	○	○	○	○	○
评估帮助学生准备考试。	○	○	○	○	○	○
评估结果是可靠的。	○	○	○	○	○	○
因为评估有误差，评估结果应审慎使用。	○	○	○	○	○	○
评估结果是比较准确的。	○	○	○	○	○	○
评估结果可用来评价教师的表现。	○	○	○	○	○	○
评估结果会被存档置然后被忽略。	○	○	○	○	○	○
评估帮助学生在考试中取得好分数。	○	○	○	○	○	○
评估是给学生判一个分数或一个级别。	○	○	○	○	○	○

评估是教学生考试技巧。	○	○	○	○	○	○
评估为未来升学和就业选择学生。	○	○	○	○	○	○
评估主要是让学生做卷子。	○	○	○	○	○	○
评估主导课堂教学进度。	○	○	○	○	○	○

如果以后从教学工作，你对实践下列活动的信心程度有多大？从左到右分别为：两个“不自信”选项，四个“自信”选项。[矩阵量表题] *

	1 非常不自信	2 不自信	3 有点自信	4 比较自信	5 很自信	6 非常自信
我能够保证评估活动和教学目标一致。	○	○	○	○	○	○
我能够把课堂上学到的评估知识运用到实践中。	○	○	○	○	○	○
我能够准确地分析学生的评估结果。	○	○	○	○	○	○
我能够用合适的方式将评估结果反馈给学生。	○	○	○	○	○	○
我能够在教学中开展形	○	○	○	○	○	○

成性评估。						
我能够制定合理的评估标准。	○	○	○	○	○	○
我能够在教学中运用多种评估方法。	○	○	○	○	○	○
我能够为学生选择合适的测试题。	○	○	○	○	○	○

从事教学活动时，你会如何开展评估活动？从左到右分别为：两个“不可能“选项，四个“可能”选项。[矩阵量表题] *

	1 非常不可能	2 不大可能	3 有一点可能	4 有可能	5 很有可能	6 非常有可能
给学生反馈时，我指出他们需要提高的地方。	○	○	○	○	○	○
我和学生共享评估目标和评估标准。	○	○	○	○	○	○
我鼓励学生对自己的学习作业进行自评。	○	○	○	○	○	○
我通过观察学生课堂表现调整自己的教学。	○	○	○	○	○	○
我用档案袋记录学生的学习过程。	○	○	○	○	○	○

我检查学生是否理解和实践了我的反馈。	○	○	○	○	○	○
我用测试作为激励学生学习的主要手段。	○	○	○	○	○	○
我监督学生互评，确保他们为同伴做出有效评价。	○	○	○	○	○	○
我根据中考或高考规则开展评估活动。	○	○	○	○	○	○
我给学生反馈时会指出他们做的好和不好的地方。	○	○	○	○	○	○
我根据国家教学大纲或省市级大纲开展评估活动。	○	○	○	○	○	○
我尽一切努力帮助学生在考试中取得好成绩。	○	○	○	○	○	○
比起课堂表现，我更重视学生的分数。	○	○	○	○	○	○
我主要用自己设计的试卷。	○	○	○	○	○	○

我开展评估活动的首要考虑是让学生在考试中取得好分数。	○	○	○	○	○	○
我通过课堂提问了解学生的理解程度。	○	○	○	○	○	○
我根据学生能力设定评估标准。	○	○	○	○	○	○
我根据测试结果反思和改进自己的教学。	○	○	○	○	○	○
我根据教学目标设定评估标准。	○	○	○	○	○	○
我经常让学生进行互评活动。	○	○	○	○	○	○
我用的测试和我的教学目标是一致的。	○	○	○	○	○	○
我用课堂上的评估活动帮助学生准备考试。	○	○	○	○	○	○
我给学生的反馈主要是分数形式。	○	○	○	○	○	○
我依据学校制定的标准进行评估	○	○	○	○	○	○

我依据测试结果了解学生学了什么。	○	○	○	○	○	○

问卷结束，非常感谢！

Appendix E

Univariate Normality Check Results

Item code	Item detail	M	SD	Skew	Kurtosis
Conceptions of assessment instrument Improvement Student development (SD)					
SD1	Assessment helps students succeed in authentic/real-world experiences.	3.41	1.113	0.257	-0.066
SD2	Assessment fosters students' character.	3.18	1.195	0.498	-0.270
SD3	Assessment is used to provoke students to be interested in learning.	3.55	1.101	0.198	-0.136
SD4	Assessment stimulates students to think.	3.93	1.056	0.023	0.228
SD5	Assessment cultivates students' positive attitudes towards life.	3.62	1.132	0.240	-0.028
Improvement Helping learning					
HL1	Assessment helps students improve their learning.	3.98	1.038	0.113	0.273
HL2	Assessment determines if students meet qualification standards.	3.40	1.136	0.211	-0.064
HL3	Assessment information modifies ongoing teaching of students.	4.23	1.051	-0.101	0.177
Improvement Accuracy					
AC1	Assessment results are trustworthy.	2.96	0.934	0.452	0.444
AC2	Assessment results are sufficiently accurate.	3.09	0.966	0.288	0.137
Irrelevance					
IR1	Assessment interferes with teaching.	2.21	0.893	1.126	2.208
IR2	Assessment results are filed and ignored.	2.74	1.108	0.691	0.369
IR3	Assessment is an imprecise process.	3.03	1.066	0.577	0.353
IR4	Assessment has little impact on teaching.	2.03	0.872	1.626	4.254
IR5	Assessment forces teachers to teach in a way against their beliefs.	2.35	1.015	1.092	1.543
Accountability Teachers and school control					
TSC1	Assessment indicates how good a teacher is.	3.14	1.083	0.343	-0.050
TSC2	Assessment is an accurate indicator of a school's quality.	2.91	1.130	0.491	0.012
TSC3	Assessment measures the worth or quality of schools.	3.25	1.165	0.252	-0.179
TSC4	Assessment is used by school leaders to police what teachers do.	3.04	1.102	0.351	-0.169
TSC5	Assessment results contribute to teachers' appraisals.	3.43	1.017	0.185	0.199

Item code	Item detail	M	SD	Skew	Kurtosis
Accountability Errors					
ER1	Teachers should take into account error and measure imprecision in all assessment.	4.45	1.101	-0.324	0.221
ER2	Assessment results should be treated cautiously because of measurement error.	4.09	1.137	-0.061	-0.172
Exams					
EX1	Assessment familiarizes students with examination formats.	3.20	1.096	0.183	-0.172
EX2	Assessment prepares students for examinations.	3.32	1.072	0.046	-0.203
EX3	Assessment helps students gain good scores in examinations.	3.15	1.058	0.229	-0.100
EX4	Assessment is assigning a grade or level to students work.	2.47	1.100	0.764	0.326
EX5	Assessment teaches examination-taking techniques	2.62	1.105	0.818	0.483
EX6	Assessment selects students for future education or employment opportunities.	3.20	1.162	0.235	-0.243
EX7	Assessment is to ask students to do test papers	2.20	0.937	1.107	1.625
EX8	Assessment sets the schedule or timetable for classes.	2.80	1.137	0.588	0.001
Self-efficacy of assessment instrument					
Self-efficacy					
SE1	I'm able to align my assessment activities with my teaching goals.	3.41	0.966	0.356	-0.007
SE2	I'm able to apply what I've learned in practical teaching.	3.60	0.937	0.240	0.274
SE3	I'm able to interpret assessment results appropriately.	3.40	0.948	0.294	0.073
SE4	I'm able to provide students with assessment results in a proper way.	3.69	0.926	0.304	0.221
SE5	I'm able to carry out formative assessment activities in my teaching.	3.54	0.965	0.302	0.184
SE6	I'm able to come up with appropriate assessment standards.	3.47	0.945	0.219	0.142
SE7	I'm able to use a variety of assessment methods in teaching.	3.63	0.980	0.213	0.058
SE8	I'm able to design the right tests for my students.	3.62	0.928	0.327	0.351
Self-reported practice of assessment instrument					
Formative practices					
FP1	I point out areas for improvement to students when I give them feedback.	4.57	0.983	-0.454	0.360
FP2	I share learning goals and success criteria with students to help their learning.	4.31	1.008	-0.263	-0.021
FP3	I encourage and value students’ own evaluation of their work.	4.48	0.973	-0.276	0.273
FP4	I observe students’ classroom performance to adjust my teaching.	4.59	0.990	-0.322	-0.047
FP5	I use portfolio assessment to record students’ learning process.	3.90	1.098	-0.116	-0.200
FP6	I check that students understand and use my	4.33	0.972	-0.230	0.313

Item code	Item detail	M	SD	Skew	Kurtosis
	feedback.				
FP7	I emphasize constructive comments over giving scores or grades.	4.49	1.022	-0.375	0.113
FP8	I ask questions in class to check students' understanding.	4.46	0.984	-0.302	0.132
FP9	I monitor student peer assessments to make sure they give each other useful comments.	4.14	1.001	-0.243	0.189
FP10	I assess according to my teaching objectives.	4.28	0.978	-0.245	0.159
FP11	I regularly ask students to give feedback on each other's work.	4.10	1.040	-0.091	-0.158
Testing practices					
TP1	I prefer getting tests from an external source rather than writing them myself.	4.22	0.976	-0.272	0.444
TP2	I use tests as a major tool to motivate students to learn.	3.73	1.101	-0.034	-0.299
TP3	I use tests designed primarily by myself.	3.42	1.057	0.148	-0.208
TP4	I value students' test results more than their classroom performance.	2.81	1.122	0.505	-0.001
TP5	I look at test results to reflect on and improve my teaching.	4.61	1.023	-0.450	0.147
TP6	The major form of feedback I give students is test results.	3.15	1.086	0.201	-0.275
TP7	I depend on test results to know what students have learned.	4.29	0.949	-0.261	0.488
TP8	My tests are aligned with teaching goals.	4.18	0.957	-0.145	0.174
Exam preparation					
EP1	I assess according to national or provincial standards or norms.	4.15	1.008	-0.225	0.120
EP2	I try my best to help students get high test scores.	4.47	1.062	-0.321	-0.131
EP3	The priority of my assessment activities is to help students get good results in their examinations.	3.48	1.104	0.030	-0.169
EP4	I assess according to *zhongkao/gaokao* rules.	3.85	1.075	-0.132	-0.033
EP5	I use classroom assessment to helps student prepare for tests.	4.02	1.002	-0.129	0.099
EP6	I regularly use practice tests with students to help them prepare for final exams.	3.96	0.981	-0.153	0.274

Appendix F

Deleted Item Details

Item code	statement	**MI**
Deleted items in C-TCoA model modification		
EX7	Assessment is to ask students to do test papers.	241.116
EX4	Assessment is assigning a grade or level to students work.	182.343
EX5	Assessment teaches examination-taking techniques	125.051
HL2	Assessment determines if students meet qualification standards.	93.194
TSC4	Assessment is used by school leaders to police what teachers do.	62.434
Deleted items in SEoA model modification		
SE1	I'm able to align my assessment activities with my teaching goals.	70.02
Deleted items in PoA model modification		
		SRW
TP4	I value students' test results more than their classroom performance.	0.11
TP6	The major form of feedback I give students is test results.	0.15
		MI
TP2	I use tests as a major tool to motivate students to learn.	211.387
TP5	I look at test results to reflect on and improve my teaching.	213.043
EP2	I try my best to help students get high test scores.	103.593
EP3	The priority of my assessment activities is to help students get good results in their examinations.	101.103
FP5	I use portfolio assessment to record students' learning process.	99.346
FP9	I monitor student peer assessments to make sure they give each other useful comments.	90.095
FP10	I ask questions in class to check students' understanding.	82.155

Note. MI=modification indices, SRW=standardized regression weights.

Appendix G

Model Fit Indices for CFA Models Modification Process

T-CoA Model Modification							
Model fit indices							
Model	χ2	χ2/df (*p*)	CFI	RMSEA	SRMR	Gamma hat	AIC
Model A (1st-order, full item)	1785.499	4. 650 (0.000)	0.859	0.059	0.0694	0.92	1947.499
Model A1 (Model A, no outliers)	1579.054	4.112 (0.000)	0.869	0.057	0.0681	0.92	1741.064
Model B (2nd-order, full item)	1912.711	4. 879 (0.001)	0.847	0.061	0.0762	0.91	2236.409
Model B1 (Model B, no outliers)	1882.440	4.754 (0.001)	0.837	0.063	0.0767	0.90	2020.440
Model C Model A-EX7	1583.053	4.447 (0.000)	0.837	0.057	0.0637	0.93	1741.054
Model C1 (Model C, no outliers)	1419.685	3.988 (0.000)	0.880	0.056	0.0628	0.93	1577.685
Model D (Model C-EX4)	1369.396	4. 162 (0.000)	0.888	0.055	0.0583	0.93	1523.396
Model D1 (Model D, no outliers)	1226.956	3.729 (0.000)	0.895	0.053	0.0576	0.94	1380.956
Model E (Model D-EX5)	1207.405	3.985 (0.000)	0.898	0.053	0.0556	0.94	1357.405
Model E1 (Model E, no outliers)	1114.325	3.678 (0.000)	0.901	0.053	0.0553	0.94	1264.365
Model F (Model E-HL2)	1069.758	3.848 (0.000)	0.907	0.052	0.0531	0.95	1220.423
Model F1 (Model F1, no outliers)	966.941	3.478 (0.000)	0.911	0.051	0.0530	0.95	1112.941
Model G (Model F-TSC4)	916.641	3.609 (0.000)	0.919	0.050	0.0479	0.95	1058.641
Model G1 (Model G , no outliers)	862.342	3.395 (0.000)	0.920	0.050	0.0486	0.95	1004.342
SEoA Model Modification							
Model fit indices							
Model	χ2	χ2/df (p)	CFI	RMSEA	SRMR	Gamma hat	AIC
Model SA (full item)	115.793	5.790 (0.000)	0.982	0.067	0.0220	0.98	163.793
Model SA1 (full item, no outliers)	78.108	3.905 (0.000)	0.988	0.056	0.0190	0.98	126.108
Model SA-SE1	52.814	3.772 (0.000)	0.991	0.051	0.0160	0.99	80.814
Model SA1-SE1 (Model SA-SE1, no outliers)	40.184	2.870 (0.000)	0.993	0.045	0.0150	0.99	68.184

PoA Model Modification							
Model fit indices							
Model	χ2	χ2/df (p)	CFI	RMSEA	SRMR	Gamma hat	AIC
Model a (full items)	2795.133	8.041 (0.000)	0.860	0.082	0.0760	0.89	2299.133
Model a1 (Model a, no outliers)	2052.943	7.520 (0.000)	0.862	0.076	0.0761	0.89	2156.845
Model b (Model a-TP4)	1658.679	6.635 (0.000)	0.894	0.073	0.0612	0.90	1758.679
Model b1 (Model b, no outliers)	1616.267	6.465 (0.000)	0.891	0.076	0.0625	0.90	1716.267
Model c (Model b-TP6)	1320.279	5.791 (0.000)	0.915	0.067	0.0510	0.92	1416.279
Model c1 (Model c, no outliers)	1301.033	5.706 (0.000)	0.912	0.071	0.0518	0.92	1397.033
Model d (Model c-TP2)	1114.250	5.383 (0.000)	0.924	0.064	0.0485	0.93	1206.205
Model d1 (Model d, no outliers)	1085.049	5.242 (0.000)	0.922	0.067	0.0493	0.93	1177.049
Model e (Model d- TP5)	934.758	4.999 (0.000)	0.935	0.061	0.0439	0.93	1022.758
Model e1 (Model e, no outliers)	909.658	4.864 (0.000)	0.934	0.064	0.0449	0.94	997.628
Model f (Model e-EP2)	812.054	4.834 (0.000)	0.940	0.060	0.0439	0.94	896.054
Model f1 (Model f, no outliers)	789.126	4.697 (0.000)	0.939	0.063	0.0441	0.95	873.126
Model g (Model f-EP3)	686.444	4.60 (0.000)	0.949	0.058	0.0328	0.95	768.444
Model g1 (Model g, no outliers)	654.735	4.394 (0.000)	0.949	0.060	0.0398	0.96	976.546
Model h (Model g-FP5)	581.627	4.406 (0.000)	0.955	0.057	0.0356	0.95	659.627
Model h1 (Model h, no outliers)	541.841	4.105 (0.000)	0.957	0.057	0.0371	0.95	619.841
Model i (Model h-FP9)	434.427	3.745 (0.000)	0.966	0.051	0.033	0.96	508.427
Model i1 (Model i, no outliers)	420.694	3.627 (0.000)	0.966	0.053	0.0357	0.96	494.694
Model j (Model i-FP10)	336.380	3.330 (0.000)	0.973	0.047	0.0312	0.97	406.380
Model j1 (Model j, no outliers)	332.121	3.288 (0.000)	0.972	0.049	0.0341	0.97	402.121

Note. CFI=comparative fit index; RMSEA=root mean square error of approximation; SRMR= standardized root mean residual; AIC=Akaike information criterion.