A Review and Discussion of In-service EFL Teachers' Language Assessment Literacy Level in Junior High School in China

Yannan Sun[1]
Teachers College, Columbia University

ABSTRACT

In China, English as a foreign language (EFL) teachers seem to be in a dilemma because of the discrepancy existing in classroom teaching and assessment, the new curriculum requirement, and the paucity of language assessment knowledge. To tackle these problems, there is an urgent need to evaluate EFL teachers' language assessment literacy (LAL) and mitigate the possible obstacles that EFL teachers face. Since the term assessment literacy (AL) was firstly suggested, the concept has been derived in other specific fields, such as LAL. Different theories, models, and research on LAL have been examined to study language teachers' LAL in many countries; however, LAL is still a new concept in China. Thus, the study aims to investigate and analyze the possible reasons behind the current LAL level of in-service EFL teachers in China.

Keywords*:* language assessment literacy, in-service EFL teachers' LAL level, reform and education, training and development

INTRODUCTION

High-stake exams in China, such as Zhongkao (High School Entrance Examination) has raised a high demand for the ability of English as a Foreign Language (EFL) teachers to design in-class assessment, analyze and evaluate test results from which they can receive feedback, and utilize the data to reflect and make improvements to their teaching methods and materials to ensure that students will be well prepared for those crucial exams (Bachman & Palmer, 1996; Tan & Ng, 2018; Zhang & Yan, 2018). This is because the test results of the large-scale examinations generally make huge impacts on school administration (Fan & Jin, 2013). Students' test results of Zhongkao, for instance, has become one of the most important factors used to indicate if an EFL teacher could impart knowledge efficiently and effectively to their students in order for them to obtain high marks on the Zhongkao English exam (Popham, 2006). Obtaining a high mark on the exam is the ultimate goal for students. (Hidri, 2020). The pressure of helping

[1] Yannan Sun received her MA in TESOL from Teachers College, Columbia University. She is currently an EFL curriculum developer at Agnaya English Institution. Her research interests include language assessment literacy, second language acquisition and second language learning and teaching. Correspondence should be sent to Email: ys3300@tc.columbia.edu

students achieve high marks in large-scale examinations has made teachers spend a large amount of time on creating assessment practices in which contents and formatting are similar to high-stake examinations. (Cheng et al., 2004). The practices of assessment conform to the concept of Assessment Literacy (AL), which requires educators to be able to distinguish a high-quality assessment from a low-quality assessment so that they can interpret their students' test results and make inferences of the students' strengths and weaknesses accurately (Bachman & Palmer, 1996; Kunnan, 2004; Stiggins, 1991; Stiggins, 1995).

Developed from AL which puts more emphasis on assessment competencies, the concept of language assessment literacy (LAL) is quite similar to that of AL. As a branch of AL, not only does LAL stipulate assessment competency, it calls for "language specific competencies" (Inbar-Lourie, 2008, p. 389), highlighting the language component that consists of "knowledge of language constructs or language assessment" (Stabler-Havener, 2018, p. 3). With the theory that has been refined continuously, LAL is widely defined through social and/or contextual factors (Coombe et al., 2020; Kremmel & Harding, 2019; Taylor, 2013).

In the past few decades, most of the research that relates to LAL tried to conceptualize what components LAL should contain and underpin the reason why classroom teachers should have knowledge of assessment skills. Although many scholars have adopted and further developed the three dimensions of LAL competencies theory which consisted of the rationale, traits, and the process of assessment (Davies, 2008; Inbar-Lourie, 2008), LAL still seems to be an ongoing topic that has been redefined and reconstructed by many others (Giraldo, 2019; Taylor, 2013; Yan et al., 2018). To illustrate, Popham (2009) suggested 13 target skills and knowledge which targeted teachers' LAL development. Later, Fulcher (2012) further expanded the definition of LAL to three layers, including contexts, principles, and practices. Pill and Harding's (2013) theoretical model advocated that LAL consisted of five proficiency levels and that LAL should be seen as a continuum, which was later modified and expanded to Taylor's (2013) profile model. According to the models, empirical research that has been conducted with the purpose of measuring classroom teachers' LAL level showed unsatisfactory results (Kremmel & Harding, 2019; Xie & Tan, 2019).

In China, most classroom teaching was teacher-centered; however, a shift from teacher-centered to student-centered approach was advocated by the Ministry of Education of the People's Republic of China (MOE, 2001) because of the guideline that called for education for well-rounded development, aiming to improve students' communicative competence and create an interactive learning environment. Nonetheless, large-scale examinations, such as Zhongkao still focuses on assessing grammatical knowledge, impeding the way teachers impart knowledge in classrooms (Wu, 2015; Zhang & Liu, 2013). To strengthen the implementation of the guidelines made in 2001, MOE (2019) decided to cancel the publication of the test specification of Zhongkao. EFL in-service teachers in junior high school who used to fully rely on the test specification are facing unprecedented challenges.

Till now, only a few researchers have focused on investigating in-service EFL classroom teachers' LAL level in China. The most prominent reason could be that current models and theories of LAL may be inappropriate to be fully implemented due to different learning styles and environment. Also, the data regarding how secondary schools and teachers select teaching materials to conduct the assessment in classroom is intransparent. Therefore, this review of literature, firstly, contributes to an understanding of the mainstream theoretical models of LAL by exploring their possible drawbacks. Next, the paper will discuss the potential problems of previous empirical research investigating in-service EFL teachers' LAL level in junior high schools in China. Finally, the stumbling blocks that in-service EFL teachers in China may face concerning their LAL level will be analyzed.

THE THEORETICAL MODELS OF LAL

This section introduces the influential theoretical models of LAL which includes the five-component model (Brindley, 2001), the model of *why*, *what*, and *how* (Inbar-Lourie, 2008), the model of skills, knowledge, and principles (Davies, 2008), the model of contexts, principles, and practices (Fulcher, 2012), Pill and Harding's (2013) continuum model and Taylor's (2013) profiles for different stakeholders.

The Five-Component Model

The model was developed in order to guide teachers to enhance their assessment level after a set of guidelines were initiated by the American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and the National Education Association (NEA) in 1990 which acknowledged teachers' assessment practices and levels. With the proposed standards as a basis, Brindley (2001) presented a five-component model for assessment development programs that provided more detailed explanations depending on various levels of assessment knowledge.

There were two essential core units and three optional units in this model. The first core unit views assessment in terms of social, educational, and political factors. Adjustment and different activities might be seen in different teaching situations and wider communities. The second core unit pertained to defining and describing language proficiency, which recounted the meaning of language ability. To design a test, the first step was to understand what language tests were and the procedures of assessment. Test developers should scrutinize the nature of the theoretical models of language ability. As Bachman and Palmer (1996) suggested, language ability could be defined through analyzing an appropriate way "for each particular testing situation, that is, for a specific purpose, group of test-takers, and [target language use] TLU domain" (p. 66).

Brinley (2001) categorized constructing and evaluating language tests, assessment in the language curriculum, and putting the assessment into practice as three optional components. First, teachers who needed to design their own tests and analyze the results should acquire the knowledge of test construction and validation with the help of available materials that were easily accessible. Second, the assessment was in relation to language curriculum, "emphasizing the close link between objectives and assessment" (Brindley, 2001, p.130). Third, it was necessary for teachers to map out follow up strategies for future policy development.

The Model of Why, What and How

The model developed by Inbar-Lourie in 2008 can be considered the integration and extension of Brinley's (2001) model through three dimensions related to LAL competencies, consisting of *why*, *what*, and *how*. The *why* section, which is equivalent to the first core unit, explained the reasoning and rationale behind the assessment. Since assessment is influenced by assessment culture in wider communities, for instance, social, educational, and political aspects in communities or societies (Brindley, 2001), the washback of language test results on decision-making should be foregrounded. The *what* section, being proportionate to Brindley's (2001) second core unit, sets forth to unfold the theoretical framework of language tests and assessment. It is exemplary that the assessment competencies of second language assessment are taken into account since English as the world language, is learnt in many countries in the world. In addition, Inbar-Lourie (2008) argued that knowledge and skills of teaching and assessment, which are parts of LAL, were equally important. A combination of the three optional units in this model can be seen in the *how* section. However, Inbar-Lourie (2008) insisted that these units were also obligatory due to the fact that language assessment development and analysis, and assessment in the language curriculum for both high-stake and classroom assessment were both necessary skills for classroom practitioners such as language teachers. Yet, teachers' personal inclinations and beliefs could affect the extent to which assessment knowledge is gained, which is another fact worthy of mention.

The Model of Skills, Knowledge and Principles

Davies (2008) reviewed language testing textbooks and discovered two trends that occurred in the past 50 years which formed his *skills, knowledge, and principles* model. The first trend was an expansion of language assessment knowledge to skills, revealing an approach of skills and knowledge. Starting from Lado (1961), which highlighted the importance for language teachers to have knowledge of language and language assessment, it offered the fundamental resources for teachers to develop language and language assessment knowledge. Since then, more practical textbooks that connect language assessment skills and knowledge were published. Bachman and Palmer's book (1996) which demonstrates the methods of test design, result analysis, statistics, and computer programs are a good case in point. Therefore, as Davies (2008) stated, knowledge serves as the background of language assessment and language description, and context settings, and skills, on the other hand, offers the assessment methods, such as item writing, statistics, test analysis, and corresponding computer programs. Principles, as the third component, concerning ethics, fairness, and impact, were added to the second trend.

In this sense, Davies's (2008) model parallels Inbar-Lourie's (2008) *why, what, and how* model. The component of *skills* is equivalent to *how* since both of them urge language teachers to use tools to assess language in practice. Both *knowledge* and *what* in the two models put emphasis on the inseparable relation between language use and language assessment. Similar to the third component *principles*, the component of *why* which looks into social themes also argues the ethnic and fairness of language evaluation in different contexts.

The Model of Contexts, Principles and Practices

Fulcher conducted an online questionnaire to 278 language teachers internationally in order to find out about teachers' needs of language assessment in 2012. Based on the result, Fulcher (2012) built a model, defining LAL into three layers: contexts, principles, and practices, which can be understood as an expansion of the previous models. The first layer *practices*, which was placed at the bottom of the layer, comprised the needed knowledge, skills and abilities for language assessment practices for both classrooms based and high-stake tests. The layer in the middle was related to *principles*, focusing on processes, principles and concepts that guide assessment practices. On the top, *contexts* referred to the knowledge of historical, social, political and ethical aspects in a wider setting, including more aspects to be taken into consideration. Notably, Fulcher (2012) pointed out the gap between policy makers, who are inclined to use high-stake tests to control educational programs. To be more specific, policy makers fail to understand the change of curriculum development and the fact that classroom practices cannot be accomplished in a short period of time especially when there is a lack of available resources in retraining language teachers to the new methodology. In return, the lack of knowledge, practical tools or skills to assess and evaluate test scores may lead to language teachers' inability to properly educate their students or have any constructive inputs on the curriculum made by policy makers who lack the experience of teaching.

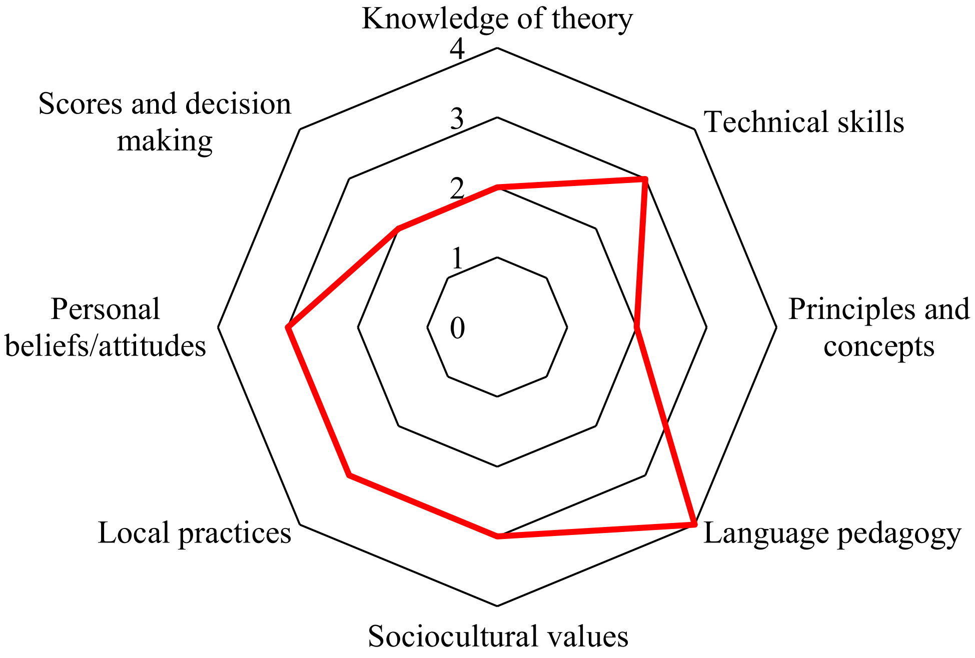
The Continuum Model for Different Stakeholders

With more attention paid to socio-economic, ethics and fairness in contexts (Brindley, 2001; Fulcher, 2012; Inbar-Lourie, 2008), more stakeholders (e.g., language assessment researchers, policy makers and students), are involved in language assessment and are in need of different levels of knowledge and skills. However, most of the research conducted, focuses on teachers' LAL levels, paying less attention to the other stakeholders. Borrowing Bybee's (1997) framework of scientific and mathematical literacy education, Pill and Harding (2013) petitioned a continuum model, intending to clarify the need of LAL of different stakeholder groups. There were five levels in total: (1) illiteracy: ignorance, (2) nominal literacy: limited understanding with possible misconceptions, (3) functional literacy: understanding basic concepts appropriately, (4) procedural and conceptual literacy: understanding and practice of key concepts, and (5) multidimensional literacy: beyond ordinary concepts and knowledge.

The Model of LAL Stakeholders' Profiles

In agreement with Pill and Harding (2013), Taylor (2013) pinpointed the necessity to contemplate and differentiate corresponding test-takers so that it would be possible to investigate the desired levels for each stakeholder group insomuch as "different stakeholder groups need different levels of assessment literacy according to their specific roles and responsibilities" (p. 408). She thereupon proposed four LAL profiles for test makers, classroom teachers, university administrators, and professional language testers. A new model whose values ranged from 0-4 (0 refers to the lowest level, namely literate, while 4 means the highest level, namely multidimensional literacy) was built based on Pill and Harding's (2013) continuum model with eight dimensions adopted from a range of papers in light of LAL, including knowledge of theory, technical skills, principles and concepts, language pedagogy, sociocultural values, local practices, personal beliefs/attitudes, and scores and decision making, which can also be found in the models mentioned above. Observing the hypothesized profiles, language pedagogy is most likely to be the most crucial aspect for classroom teachers 4, followed by local practices, sociocultural values, personal beliefs/attitudes, and technical skills 3, indicating that teachers should always be aware of the test environment and any policy changes regarding high-stake examinations in one particular context (see Figure 1). This echoes what Fulcher (2012) claimed large scale tests "are used in political systems to manipulate the behavior of teachers and hold them accountable for much wider policy goals" (p. 1).

FIGURE 1
LAL Profile for Classroom Teachers (Taylor, 2013, p. 410)



Kremmel and Harding (2019) designed an online survey to investigate the LAL level of different stakeholders based on Taylor's (2013) profile model. What they discovered was that language teachers thought they should have a balanced role in all the requirements related to LAL and that the average for most of the dimensions was a 3, higher than what Taylor (2013) hypothesized. Noteworthily, Kremmel and Harding (2019) argued that there should be nine components to LAL rather than eight, adding developing and administering language assessments as one component.

Possible Barriers of the Models

All models attempted to define what LAL is and what it means to language teachers. Undoubtedly, all researchers made significant contributions to the further development of LAL, the models of fundamental principles and the concept of LAL to stakeholders from different groups. Some drawbacks to these models can be seen at the same time, however. The most obvious problem would be the absence of a finalized conception and construct of LAL (Inbar-Lourie, 2013), making LAL mysterious and complex that it is difficult to comprehend. Confusions and misunderstanding may arise if LAL is not viewed meticulously because of the remarkably similar components, which are named differently, but with the same set of terminologies. For instance, the component *why* in Inbar-Lourie (2008) is changed to *principles* in Davies (2008). Bewilderingly, *principles* in later Fulcher (2012) refers to the guidance for assessment in practice rather than the social role of assessment such as ethics.

Another problem could relate to the framework of LAL that probably will continuously evolve according to new curriculum and policy. Hence, transparency between policy makers and teachers should be guaranteed. Otherwise, difficulties can arise when it comes to incorporating assessment with curriculum and criteria, as Coombe et al. (2020) argued. If this is the case, long-term training is necessary in order to help teachers be familiar with any updated knowledge on assessment and policy in time. However, in many countries, such as China, communications between stakeholder groups might not be enough.

None of the models conceptualized LAL with enough straightforward and sufficient information and details, enabling teachers to know what exactly they need to learn. To fill this gap, Popham (2009) advocated 13 suggestions with regard to knowledge and skills of assessment. What makes these suggestions valuable are the reflection on students' perceptions and attitudes from assessment. To take one example, teachers' ability to use assessment results as evidence to measure "students' attitudes, interests and values" (Popham, 2009, p. 9) is essential since these factors may bring a negative long-lasting impact on students' future learning. Besides, he argued that teachers should consider using a wide range of assessment methods (i.e., portfolio, and peer- and self-assessments) rather than merely applying selected response questions such as multiple choice and true or false questions to build a connection between classroom assessment and large-scale tests. This point he made may be quite insightful. Though Popham (2009) did provide more details and instructions for teachers, the target of the suggestions was assessment, not language assessment. Problems related to what assessment methods should be considered in specific and the level to which teachers need to understand assessment, use statistical tools to design and select appropriate tests remained unclear.

Regarding the burning issues of LAL and language teachers' LAL level, a model integrated with a specific culture or social context could be beneficial because of the wide variety of policies, teachers' training programs, testing culture, etc. from country to country. Therefore, investigations of language teachers' LAL level, data collection, and analyses should be conducted within one nation in order to build a more detailed and useful guidance of LAL development.

IN-SERVICE EFL TEACHERS' LAL LEVEL IN SECONDARY SCHOOL IN CHINA

Even though there are not many studies about in-service EFL teachers' LAL level in China, the flaws of the models can be observed in existing ones, especially in relation with socio-cultural issues. Given the importance of language teachers' LAL development, this section discusses and analyzes the downsides of the chosen empirical studies that analyzed the in-service EFL teachers' LAL level in junior high schools in China. Next, some of the prominent factors contributing to measuring language teachers' LAL level is analyzed to conclude the section.

The Research on Measuring In-service Teachers' LAL Level and Their Problems

Zhang and Yan's (2018) research looked into in-service EFL teachers' LAL level especially, with their ability to determine whether test items in an end-of-year English tests were of high standards and whether the test items were at appropriate difficulty levels for their students. There were 378 in-service EFL teachers from one city in the northern part of China participating in the research. Contrary to the assumption that in-service EFL teachers in China lack the ability to develop good test items or the claim that they are almost illiterate in terms of assessment knowledge and skills (Huang & He, 2016; Huang and Jiang, 2020; Sun & Shines, 2020), Zhang and Yan (2018) revealed that EFL teachers were able to come up with satisfactory psychometric quality items on average since the test items showed high internal consistency. In addition, in-service teachers were capable of perceiving the difficulty level of the test items that hinged on their students' proficiency level accurately even without LAL professional training. When the test items were analyzed by the two researchers, they discovered some problems that were not detected by the participants. Therefore, they acknowledged the need for LAL development and training for in-service teachers because "this lack of training in item analysis might have prevented these secondary teachers from identifying problematic items for revision and making appropriate decisions and inferences in the use of test items and score" (Zhang & Yan, 2018, p. 43).

A few problems exist in the research, however. First, the conclusion drawn of in-service EFL teachers' LAL level from one single test might be insufficient to make inferences on the overall quality of the tests. Besides, the participants of the study were all from one city in China, which cannot be generalized to the LAL level of all the teachers. There is the possibility that the results were caused by chance, leading to inaccurate interpretations of their LAL level. Additionally, the interpretation of one of the test items might be incorrect. Zhang and Yan (2018) criticized that teachers' English language proficiency might be problematic since both item writers and the participants failed to spot out the error made in item 28, suggesting that the correct answer "open up" was a non-idiomatic expression used in the item "Young Lei said she could _______ her students' eyes to the outside world (p. 42).

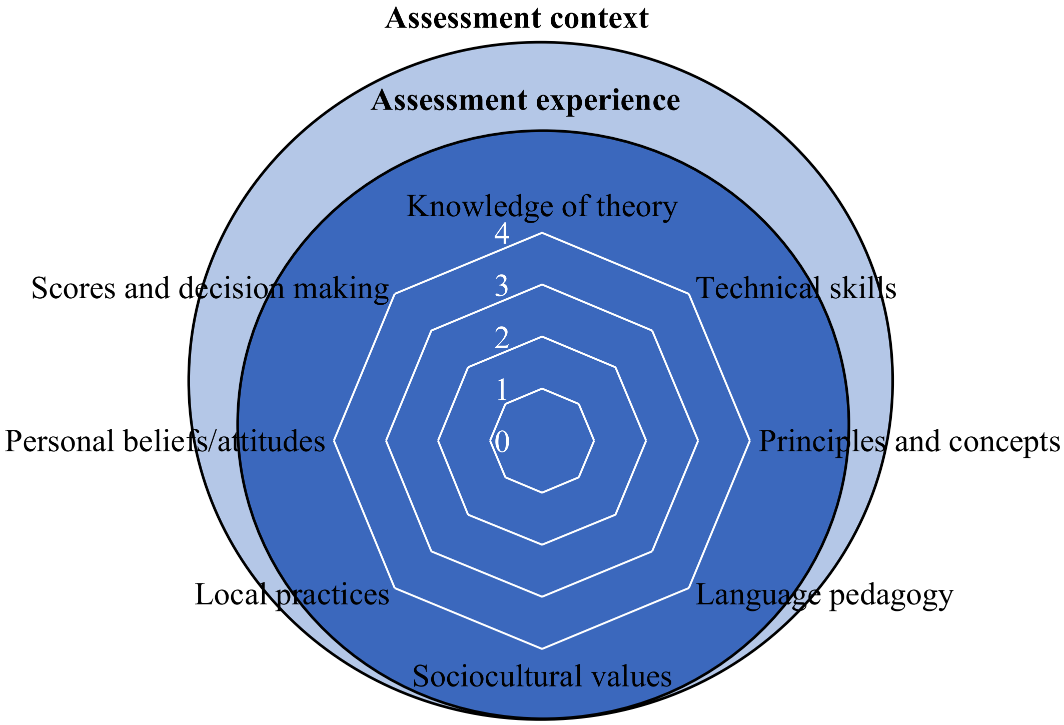
To provide a better picture of teachers' LAL level, Yan et al. (2018) interviewed three in-service EFL secondary school teachers from the same school in northern China, using Taylor's (2013) profile, and Pill and Harding's (2013) continuum models. The paper found a similar LAL profile and training need patterns as Tayor (2013) claimed, reflecting that EFL teachers' LAL level was not as low as predicted, as Zhang and Yan (2018) stated. Though the three attendants in the study did not receive LAL professional training, they developed their assessment knowledge and skills through two mediators: assessment context and assessment experience (see Figure 2). Yan et al., (2018) opined:

> That is, changes in the assessment context can alter teachers' assessment practice routines, thereby influencing their interests and likelihood in developing specific assessment knowledge and skills...teachers can cultivate the assessment culture in the local context, which can help expand their assessment practice, ultimately leading to continued development of LAL (p. 166).

Nonetheless, assessment context and experience as two additional layers added to Taylor's (2013) profile with a special emphasis on how teachers' LAL developed in the context of China was created.

FIGURE 2

Mediation of Contextual and Experiential Factors on Teachers' LAL Development (Yan et al., 2018, p. 166)



The limitation of the study is that the three participants were acquaintances of the researchers. Apart from the six years of teaching experience, the three participants also had experience of taking part in school-based and district/city-wide final English tests design. However, a large number of EFL teachers in China are novice teachers who do not have extensive teaching experience to remedy their LAL level. Hence, the results do not represent the overall teachers' LAL level in other regions despite the researchers choosing the experienced participants deliberately to find out about the potential factors related to LAL development.

In 2019, Lan and Fan designed a questionnaire to investigate 344 EFL teachers' current and expected LAL levels in secondary schools in service from different regions who had various degrees of professional training through a combined model integrating Pill and Harding's (2013) continuum, Vogt and Tsagaris' (2014), and Taylor's (2013) profile models. Five areas were taken into account, including technical skills, scores and decision making, language pedagogy, knowledge of theory and principles and concepts. The result showed in-service teachers' LAL level was nearly at a functional level, around 2 (ranging from 0-4) for each category. The categories of principles and concepts, knowledge of theory and scores and decision making about language assessment indeed met what Taylor (2013) hypothesized, while the scores of technical skills and language pedagogy were lower. According to Taylor's (2013) profile, language pedagogy should be the highest (4); however, the result was just above 2, slightly higher than the others, suggesting that EFL teachers may impart knowledge ineffectively. Similarly, the technical skills may indicate that in-service teachers were not qualified enough to handle high quality assessments in classrooms.

The biggest setback with the research is similar to that of Kremmel and Harding (2019)'s study which was that individuals' perceptions of each level might not be the same. Since the participants had various levels of LAL training, they may interpret items differently as well. As a result, the scales that the participants used are most likely to have standard errors of measurement, leading to inaccurate interpretations of the result. It is interesting that the data for item 102: using ready-made tests was higher than that of item 101: preparing classroom tests (teachers' perceived LAL level), although the participants expected to design their own tests. This may indicate that in-service EFL teachers in junior high schools in China rely more on pre-made tests and lack the ability to design tests on their own. Over and above that, the participants' knowledge of using portfolios as assessment ranged only between 1 and 2, the lowest among all the items, "close to the level of nominal literacy" (Lan & Fan, 2019, p. 118). Although the data was expected to have higher portfolios use (just below 2.8), it was still lower than the desire to develop skills in using self-made and pre-made tests by others, supporting Huang and Jiang's (2020) argument that secondary English teachers are generally aware of the value of authentic assessment, such as portfolios, but rarely use them because of factors, including "high-stakes testing culture, limited resources, teachers' assessment literacy and misconceptions about students" (p.12)

To summarize, research evaluating in-service EFL teachers' LAL level in China exhibits inconsistent findings. Even so, the studies come to the agreement that the LAL level of in-service EFL teachers is comparatively low, and it is necessary to study the essential reasons behind the result to gear toward advancing teachers LAL level in the future.

Factors Contributing to the Low LAL

Examination Oriented Environment

Although the MOE (2001) requires in-service EFL teachers to abandon the traditional assessment methods and the teaching approach that targets on grammar, teachers may still be reluctant to make a change in junior high schools in China. The competitive environment forces both in-service teachers and students to devote themselves to preparing for Zhongkao. Since grammatical knowledge is still the focus of Zhongkao, students are being evaluated with traditional language assessment rather than authentic assessment that could help them be able to deal with language use after leaving school (Huang & Jiang, 2020). According to Jin et al., (2017), classroom-based assessment has three main purposes: 1. Measuring students' learning to identify what students need to get better grades; 2. Measuring teachers' effectiveness to adjust instructional decisions; 3. Measuring students' learning outcomes compared to national curriculum and standards. In other words, the exam-oriented environment in China makes teachers generally pay less attention to training students to use language in real life and following MOE's (2001) guideline.

The Lack of Communication with Policy Makers

EFL teachers in China who face the pressure of high-stake examinations use assessment as a means to measure how well their students are prepared for the examinations. According to the reform by MOE (2001), assessment should be seen as a feedback for teachers to improve their teaching methods and efficiency to promote students' learning progress. Teachers need to educate students in an interactive way to help them improve both linguistic and communicative competence. However, in reality, classroom teaching, and assessments were still examination oriented since Zhongkao remained the same (Shao, 2015). To urge teachers to follow the guideline, MOE (2019) announced the cancellation of test specifications for junior high schools in mainland China to prevent the phenomenon of learning for tests and to improve students' ability in using English. The sudden change could be troublesome for teachers to change and find suitable learning materials and assessments for their students due to the heavy reliance on test specifications in the past. However, teachers might use the past Zhongkao paper and sample answers to teach and assess students, leading to the failure of implementing the new requirements. Therefore, policy makers should interact with classroom teachers and provide a general outline for teachers to follow to enforce the new national curriculum.

Teachers' Personal Beliefs

As the models discussed above suggested, teachers' personal beliefs regarding language assessment can be another factor causing the comparatively low LAL level of EFL teachers in China and hindering LAL development in the future. Suggested by Zhang and Liu (2013), teachers' personal beliefs are shaped and influenced by multiple strands. The study pinpointed that teachers' beliefs influence how they teach. In-service EFL teachers in junior high schools have gradually accepted the constructivist ideas which put stress on enhancing students' communicative abilities and participation in class because of the new national curriculum. Having said that, many teachers still keep the traditional views on teaching pedagogy, such as memorizing vocabulary, sentence structures, and reading passages due to several contextual factors, such as curriculum reform, large-scale examinations, traditional cultural values, and school types. In spite of the discussed factors related to high-stake examinations and curriculum reform, school types are also influential. Zhang and Liu's (2013) research brought attention to the impacts of the school environment and culture as more prestigious schools may better facilitate constructivist beliefs of teachers, showing that local practices and socio-cultural values play a role in affecting language teachers' LAL level and development.

Insufficient Training

Insufficient training can also account for in-service EFL teachers' LAL level in China because of the absence of language assessment knowledge requirement in teacher's qualification test and LAL training for in-service teachers. For people who wish to teach in public schools, they must pass a two-part examination to get the teacher's qualification certificate. The first part is a paper-pencil test that measures candidates' comprehensive quality, pedagogical knowledge, teaching ability, and language ability (grammar, writing, and reading in particular). The other part is an interview in which candidates give a demo class to the examiners for approximately 15 to 20 minutes (National Education Examination Authority [NEEA], 2012). The whole test does not put much emphasis on the candidates' language assessment ability, though. Apart from this, in-service EFL teachers in China may receive inadequate professional training. He (2019) reported that there were only 39% of teachers in the survey who had language assessment training experience, and teachers in general needed to rely on their personal teaching experience to conduct tests. The neglect of training teachers' LAL can also be sensed in one of the cities in China. Headmasters and in-service teachers from elementary and middle schools combined together, only 30 people had the opportunity to attend the training that somewhat touched on language assessment (MOE, 2020). Admittedly, teachers' experience can compensate for the insufficient training as was suggested by Yan et al. (2018); however, LAL training is still crucial, especially the EFL teachers who have no teaching experience to rely on. This might be a plausible explanation of why it is a norm that Chinese parents and students prefer senior teachers and are unwilling to learn with young teachers.

CONCLUSION

Bearing in mind the examination-oriented environment and regarding the continuously changing policies of education, the primary objectives of this review were to examine in-service EFL teachers' LAL level in junior high school in China and analyze the possible reasons behind the current level by reviewing LAL models and their empirical studies in China. To understand what LAL and its models were, the review started with an introduction of the background of the high-stake examination Zhongkao and its impact on both micro and macro levels in China. With the intention of assisting their students to get a perfect grade in Zhongkao, in-service EFL teachers have been using Zhongkao exam questions as the major assessment to evaluate their students' learning process and outcome. Therefore, the quality of the assessment that language teachers use is put under the spotlight, especially after the cancelation of test specifications of Zhongkao. Thus, a summary of the influential models of LAL developed over the past decades was discussed to clarify what LAL is because of their complexity and differences in defining certain terminologies. Next, this paper looked into a few empirical studies conducted in China since socio-culture and testing environments both have huge impacts on language teachers' LAL level and development. After reviewing and delving into the potential problems that may exist in each empirical study, four possible factors that could cause the unsatisfactory LAL level of EFL teachers in-service in China were investigated.

Limitations

There are a few limitations in this review of literature. The foremost one is the limited number of both nonempirical and empirical research. Also, the analysis of some of the research could be inaccurate. To illustrate, Yan et al.'s (2018) research may only represent a small group of individuals' points of view and LAL level because of the small sample size. Thus, it is reckless to use the model to conclude EFL teachers' LAL level national-wide in China. This is also related to another limitation of this review. Whether it is a high-stake examination or not, classroom assessment and Zhongkao have different qualities and versions. In other words, depending on the different curriculum developed by schools and the MOE of each province in China, there might be a gap between one area to another. Hence, it would be impossible to generalize EFL teachers' LAL level of one particular area to others. Another limitation would be the restricted analyses in terms of the factors contributing to the low teachers' LAL level. Factors such as teachers' ability in understanding and applying language assessment should be considered. The models discussed in this paper and the data reported about test validity for instance are most likely to be hard for non-linguists or people who have never had assessment training. Being viewed as a practical book for non-experienced assessors, Bachman and Palmer's textbook published in 1996 could still be incomprehensible (Taylor, 2013).

Implications

The overview of the models of LAL and their empirical studies shed light on building a better understanding of the theoretical models of LAL and interpreting studies of in-service teachers' LAL level in China by reflecting on the potential issues of the papers. Notwithstanding, many outstanding definitions and models of LAL that have been presented, LAL is still under construction because of various discrepancies in the communication between language teachers and policy makers, teachers' own beliefs, various contexts related to socio-culture, and insufficient teacher training in different areas. The intricacy of a unified model that can be applied in general and be used to measure language teachers' LAL level is problematic. Yet, regardless of the challenges, LAL is essential to improve one nation's language education. Therefore, it is necessary to have more research exploring LAL with the target in one particular context to improve language teachers' LAL level with detailed instructions.

Returning to the major concern of the paper that was about in-service EFL junior high school teachers' LAL level in China, there are recommendations for stakeholders. Policy makers in China need to consult with in-service EFL teachers in advance before any decisions are made at large-scale tests. No one can understand what students learn and how well they learnt than in-service teachers. Teachers' pedagogy and teaching materials can influence students' learning progress and outcomes to make sure large-scale tests assess what students are being taught in class. Thus, policy makers should communicate and work with teachers together to make changes to "curriculum development …[and] classroom practices" (Fulcher, 2012, p. 1). Any sudden change in curriculum, classroom practices, and high-stake tests may cause negative washback which may influence administration's decisions if teachers misunderstand reforms.

Reform towards cultivating students' ability in using English should be encouraged, but also reform should fully consider the long history of the rigid examination-based environment and Chinese cultural values, which may not be easily altered. For example, although MOE has been emphasizing the importance of education for well-rounded development, there are no significant differences in the way teachers teach and assess students because Zhongkao remains the same. In contrast, Gaokao (College Entrance Examination), which just completed its reform in 2020 can be an example. The new Gaokao focuses more on assessing students' ability in using English so that inferences of the students' English language proficiency can be made (Cheng, 2008). To illustrate, the tasks in the writing section not only aim to test students' language ability but also their coherence and logic, demanding classroom teachers to impart knowledge and develop assessment focusing on students' integrated skills (e.g., reading, writing, and grammatical knowledge). Additionally, authentic assessment materials are used by teachers in classrooms in terms of being in line with the new type of continuation task since it requires test-takers to comprehend one reading passage and continue to write the story based on their understanding (Wang & Zhang, 2018). In this case, the argument that language teachers' personal beliefs are influenced by their past learning experience rather than "assessment theories or policy requirements" (Coombe et al., 2020, p. 10) may not be persuasive since the large-scale test left teachers no other options but to change.

Furthermore, in-service EFL teachers in China should be provided with professional training to develop their LAL knowledge and skills for the long term. For EFL teachers in China who have relatively low LAL level at the current stage, the quality of Zhongkao and in-class assessment might have poor quality. Teachers who heavily rely on past sample papers and answers when designing in-class tests should consider if the adopted assessment is valid and suitable to use for their students. To change the situation, organized LAL training designed for in-service EFL teachers is needed.

REFERENCES

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford University Press.

Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, McNamara, T., O'Loughlin, K., (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 126–136). Cambridge University Press.

Bybee, R. W. (1997). *Achieving scientific literacy: From purposes to practices.* Heinemann.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, *25*(1), 15–37. https://doi.org/10.1177/0265532207083743

Cheng, L., Rogers, T., & Hu, H. (2004). ESL/EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing*, *21*(3), 360–389. https://doi.org/10.1191/0265532204lt288oa

Coombe, C., Vafadar, H., & Mohebbi, H. (2020). Language assessment literacy: What do we need to learn, unlearn, and relearn? *Language Testing in Asia*, *10*(3), 1–16. https://doi.org/10.1186/s40468-020-00101-6

Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, *25*(3), 327–347. https://doi.org/10.1177/0265532208090156

Fan, J., & Jin, Y. (2013). A survey of English language testing practice in China: The case of six examination boards. *Language Testing in Asia*, *3*(1), 1–16. https://doi.org/10.1186/2229044337

Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, *9*(2), 113–132. https://doi.org/10.1080/15434303.2011.642041

Giraldo, F. (2019). A post-positivist and interpretive approach to researching teachers' language assessment literacy. *Profile: Issues in Teachers' Professional Development*, *22*(1), 189–200. https://doi.org/10.15446/profile.v22n1.78188

He, H. (2019). *A comparative study of language assessment literacy between pre-service and in-service teachers* (dissertation). Guangdong University of Foreign Studies.

Hidri, S. (2020). Language assessment literacy: Where to go. In S. Hidri (Ed.), *Perspectives on language assessment literacy: Challenges for improved student learning* (pp. 1–10). Routledge.

Huang, J., & He, Z. (2016). Exploring assessment literacy. *Higher Education of Social Science*, *11*(2), 18–27. https://doi.org/10.3968/8727

Huang, R., & Jiang, L. (2020). Authentic assessment in Chinese secondary English classrooms: Teachers' perception and practice. *Educational Studies*, *47*(6), 633–646. http://doi.org/10.1080/03055698.2020.1719387

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing, 25*(3), 385–402. https://doi.org/10.1177/0265532208090158

Inbar-Lourie, O. (2013). Language assessment literacy. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp.1–9). Blackwell Publishing Ltd.

Jin, Y., Wu, Z., Alderson, C., & Song. (2017). Developing the China standards of English: Challenges at macropolitical and micropolitical levels. *Language Testing in Asia*, *7*(1), 1–19. https://doi.org/10.1186/s40468-017-0032-5

Kremmel, B., & Harding, L. (2019). Towards a comprehensive, empirical model of language assessment literacy across stakeholder groups: Developing the language assessment literacy survey. *Language Assessment Quarterly, 17*(1), 100–120. http://doi.org/10.1080/15434303.2019.1674855

Kunnan, A. J. (2004). Test fairness. In M. Milanovic and C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge University Press.

Lado, R. (1961). *Language testing*. Longmans.

Lan, C., & Fan, S. (2019). Developing classroom-based language assessment literacy for in-service EFL teachers: The gaps. *Studies in Education Evaluation*, *61*(3), 112–122. http://doi.org/10.1016/j.stueduc.2019.03.003

Ministry of Education of the People's Republic of China. (2001, February 6). *The important changes towards well-rounded education.* http://www.moe.gov.cn/srcsite/A08/s7056/200102/t20010206_162632.html

Ministry of Education of the People's Republic of China. (2019, November 30). *Three measurements of changing education.* http://www.moe.gov.cn/fbh/live/2019/51594/mtbd/201912/t20191203_410647.html.

Ministry of Education of the People's Republic of China. (2020, November 9). *The need of improvement of education and development in Dalian*. http://www.moe.gov.cn/jyb_xwfb/s6192/s222/moe_1768/202011/t20201109_498999.html.

National Education Examination Authority. (2012, May). *The law of teacher qualification regulation*. http://ntce.neea.edu.cn/html1/category/1511/692-1.htm

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381–402. https://doi.org/10.1177/0265532213480337

Popham, W. J. (2006). Needed: A dose of assessment literacy. *Educational Leadership, 63*(6), 84–85.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice, 48*(1), 4–11.

Shao, W. (2015). *Study on language assessment literacy training needs of pre-service and in-service middle school English teachers* (dissertation). Guangdong University of Foreign Studies.

Stabler-Havener, M. L. (2018). Defining, conceptualizing, problematizing, and assessing language teacher assessment literacy. *Working Papers in Applied Linguistics & TESOL*, *18*(1), 1–22. https://doi.org/https://doi.org/10.7916/salt.v18i1.1195

Stiggins, R.J. (1991). Assessment literacy. *The Phi Delta Kappan, 72* (7), 534–539.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *The Phi Delta Kappan, 77*(3), 238-245. http://www.jstor.org/stable/20405538

Sun, Q., & Shines, A. J. B. (2020). A meta-analysis on EFL secondary teachers' professional competences in China. *Journal of Critical Reviews*, *7*(6), 246–264.

Tan, C., & Ng, C. S. L. (2018). Assessment reform in Shanghai issues and challenges. *International Journal of Educational Reform*, *27*(3), 291–309. https://doi.org/10.1177/105678791802700304

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30(3),* 403–412. https://doi.org/10.1177/0265532213480338

Vogta, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11(4),* 374–402. https:// 10.1080/15434303.2014.960046

Wang, D., & Zhang, X. (2018). National matriculation English test in China: Its past, present and future. *Journal of Applied Linguistics and Language Research*, *5*(4), 183–199.

Wu, Y. (2015). The examination system in China: The case of Zhongkao mathematics. In S. Cho (Ed.), *Selected regular lectures from the 12th International Congress on mathematical education* (pp.897-914). Springer.

Xie, Q., & Tan, S. (2019). Preparing primary English teachers in Hong Kong: Focusing on language assessment literacy. *The Journal of Asia TEFL*, *16*(2), 653–673. http://dx.doi.org/10.18823/asiatefl.2019.16.2.14.653

Yan, X., Zhang, C., & Fan, J. J. (2018). "Assessment knowledge is important, but ...": How contextual and experiential factors mediate assessment practice and training needs of language teachers. *System*, *74*, 158–168. https://doi.org/10.1016/j.system.2018.03.003

Zhang, C., & Yan, X. (2018). Assessment literacy of secondary EFL teachers: Evidence from a regional EFL test. *Chinese Journal of Applied Linguistics*, *41*(1), 25–46. https://doi.org/10.1515/cjal-2018-0002

Zhang, F., & Liu, Y. B. (2013). A study of secondary school English teachers' beliefs in the context of curriculum reform in China. *Language Teaching Research*, *18*(2), 187–204. https://doi.org/10.1177/1362168813505940