English Language Teaching; Vol. 13, No. 9; 2020
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

A Rasch-Based Validation of ELT Certificate-LORT

Xin Qu[1]

[1] School of English Studies, Beijing International Studies University, P. R. China

Correspondence: Xin Qu, School of English Studies, Beijing International Studies University, No. 1, Nanli Community, Dingfu Town, Chaoyang District, Beijing, P. R. China.

Received: August 7, 2020 Accepted: August 24, 2020 Online Published: August 25, 2020

doi: 10.5539/elt.v13n9p94 URL: https://doi.org/10.5539/elt.v13n9p94

Abstract

The present study was executed with the purpose of validating ELT Certificate Lesson Observation and Report Task (ELTC-LORT), which was developed by China Language Assessment to certify China's EFL teachers by performance-based testing. The ELT Certificate has high-stakes considering its impacts on candidates' recruitment, ELT in China and quality of education, so it is crucially important for its validation so as to guarantee fairness and justice. The validity of task construct and rating rubric went through a process suited for many-facet Rasch measurement supplemented with qualitative interviews. Participants (N = 40) were provided with a video excerpt from a real EFL lesson, and required to deliver a report on the teacher's performance. Two raters graded the records of the candidates' reports using rating scales developed to measure EFL teacher candidates' oral English proficiency and ability to analyze and evaluate teaching. Many-facet Rasch analysis demonstrated a successful estimation, with a noticeable spread among the participants and their traits, proving the task functioned well in measuring candidates' performance and reflecting the difference of their ability. The raters were found to have good internal self-consistency, but not the same leniency. The rating scales worked well, with the average measures advancing largely in line with Rasch expectations. Semi-structured interviews as well as focus group interviews were executed to provide knowledge regarding the raters' performance levels and the functionalities of the rating scale items. The findings provide implications for further research and practice of the Certificate.

Keywords: ELT Certificate-LORT, performance assessment, many-facet Rasch measurement, validation, EFL teachers' competence

1. Introduction

1.1 Background

English language teaching certificate is an essential facilitator to measure English teachers'(Note 1) competence, qualify their teaching status, and encourage their professional development. This certification category diversifies by adopting paper-pencil tests, performance assessments, or training programs to conform to specific certificate constructs.

As an important issue, English language teaching certificates have been explored internationally in recent years. In the 1960s, UK and US authorities concerned with education started testing prospective language teachers; such tests were initiated with the direct objective of ensuring that teachers would satisfy minimum requirements stipulated for fundamental skills and subject knowledge. During 1980s and 1990s, assessment credentials have been world widely used for teacher selection and teacher development (Angrist & Guryan, 2004; Latham et al., 1999; Pulverness, 2015).

Some certificates are developed for a variety of English teaching contexts around the world by prominent organizations such as Cambridge English Language Assessment and Educational Testing Service. Others are designed to meet local requirements. The former group is extended with certificates like Certificate in Teaching English to Speakers of Other Languages (CELTA), Teaching Knowledge Test (TKT), and ELTeach. The latter comprises English Exam for Teacher Certification (EETC), Praxis-ESOL, and the Language Proficiency Assessment for Teachers (English Language) (LPATE), etc.

1.2 China's Situation

Besides being an important issue explored internationally, English language teaching certificate is also an essential requirement of China's EFL (Note 2) education in recent decades. Around 1990s, China's Ministry of Education initiated assessment credentials for teacher qualification. In 2011, standards were promulgated for school and kindergarten teacher certification. Since 2015, it has been required national widely for teacher candidates to pass the examination of teacher certification.

Given the quickly developing landscape of present English language teaching reform, primary and secondary school EFL teachers, than ever before, crave for more continuing education and sustainable development to accredit themselves as professionals. Performance-based testing and standards-based learning are an inevitable way out for their professionalization and professionalism.

Against this background, an English language teaching certificate is being born to evaluate and certify their core competencies in close relation to local teaching practice. Hence the English Language Teaching Certificate (the ELTCertificate for short), developed and designed by China Language Assessment (CLA) since 2015 as a significant instrument for China's EFL teacher certification, bears great interests in local teachers' education, recruitment, and development.

1.3 The Present Study

The objective of this study was to validate one task of the ELTCertificate: Lesson Observation and Report Task (ELTC-LORT). This task is believed to provide information regarding EFL teachers' abilities of oral reporting and evaluation of classroom ELT practice. The performance of ELTC-LORT was statistically analyzed and discussed through many-facet Rasch measurement (MFRM), which enabled consideration of assessment variables, such as rater severity and task difficulty, in estimating a candidate's underlying ability (Lynch & McNamara, 1998, p.161; Stephens, 2018). Qualitative measures of semi-structured interview and focus group interview usefully supplement and extend the MFRM analysis.

The study is guided by the following research questions:

(a) How do the MFRM analysis demonstrate an estimation and (b) Is ELTC-LORT accountable in measuring candidates' performance?

2. Literature Review

Starting from the 1950s, the development of foreign language education has enriched research that has been executed in the academic domain of EFL teacher education and evaluation; specifically, such research has evolved with studies of linguistics and applied linguistics as well as those in the field of language teaching. Academic studies in this domain have explored topics including what language teachers should be required to know, what they should know, and what they seek to know.

In the mid-20th century, researchers focused on describing the basic skills such as vocabulary and grammar that language teachers should possess. The Competency-Based Teacher Education (CBTE) movement in the 1960s provided a hierarchical classification and item-by-item description of teachers' teaching behavior. From the late 1970s to the 1980s, the development of communicative teaching required the professional knowledge and skills of language teachers to be based on meaningful language use. In the 1990s, the study of language teachers' professionalization and professionalism began to focus on teachers' cognition and explore how teachers learn to teach. Since the new century, in the perspective of teaching practice, researchers no longer regard teachers as "outsiders" who acquire and teach knowledge, but "insiders" who master and create knowledge and understand complex teaching environment.

2.1 From 1950s to 1960s

Around the Second World War, because of the effect of behavioral psychology and structural linguistics, practitioners in the domain of language teaching in the United States highly accentuated the following: teachers' speaking, listening, writing, and reading skills with respect to language; their mastery of grammar and vocabulary; and their understanding of language, culture, and pedagogical knowledge.

The knowledge base of foreign language teachers' language competence founded in this historical period was practical and easy to be standardized to a certain extent, but also brought about the problem of de-contextualization. Nevertheless, through the conceptual description of this period, the characteristics of foreign language teachers' language competence have been presented and distinguished from those of other disciplines, which lays a foundation for systematic research in this field.

"It is a period when knowledge-for-teaching tended to be defined almost exclusively as content knowledge" and

was equated to knowing the subject matter of language (Freeman et al., 2009, 81). This was noted to exert a direct influence on the judgment of foreign language teacher qualification. Echoing to the idea that teachers wouldn't be competent for teaching without knowledge of language, tests of language proficiency turned to be synonyms for foreign language teacher qualification, with paper-and-pencil mode and multiple-choice questions as primary test design.

2.2 From 1970s to 1980s

In the 1970s and 1980s, achievements in the academic field of linguistics stimulated researches of language teaching. The probe into teachers' language ability also gained inspiration and guidance in the study of linguistics and applied linguistics. *The Edinburgh Course in Applied Linguistics* and *Georgetown University Round Table on Languages and Linguistics* devoted special chapters to the study of language teaching and language teachers. Accomplished linguists and thoughtful researchers indicated readers their in-depth understanding of teachers' language competence, among them wrote Noam Chomsky, Pit Corder, Larsen-Freeman, Janice Yalden, and other authors. Researches in linguistics threw lights on new views of language ability, hence brought new requirements for research and practice of language teaching. The theory of communicative competence and its application in language teaching have revolutionized the connotation of the concept of language teachers' professional competence.

Studies of this period changed the foci from professional knowledge and skills to language teachers. Researchers also shifted from demonstrating what kind of knowledge language teachers should master to the process through which they master knowledge, the context within which they apply knowledge and the methods with which they teach language. Meanwhile, the concept of equating subject knowledge with teaching knowledge has been questioned. Shulman (1987) argued for a distinction between "pedagogical content knowledge (PCK)" and disciplinary "content knowledge". PCK integrates subject knowledge and teaching knowledge, and was regarded as foundation of language teachers' professional competence. The assessment of teachers' language proficiency was no longer a "lifelong test", but more concerned about the progress of language teachers' professional development. Portfolio became a recording and evaluating means for "language teachers to increase their professional knowledge and improve their professional skills" (Gibbs, 1983; Rutherford, 1987; Shulman, 1988; Weinberger & Didham, 1987). However, how to effectively examine language teachers' competence turned out to be a problem to be solved.

2.3 From 1990s to the New Century

If contemporary researches on language teachers' competence initially draw theoretical nourishment from linguistics and applied linguistics, the academic achievements of psycholinguistics, sociolinguistics, cultural linguistics and cognitive linguistics provide new resources for studies in this field.

"In recent years, L2 teacher education has been influenced by viewpoints drawn from sociocultural theory emphasizing that learning is a 'situated activity' that takes place in specific social contexts; teacher learning is not viewed as translating knowledge and theories into practice but as constructing new knowledge and theory through participating in specific social contexts and engaging in particular types of activities and processes" (Lave & Wenger, 1991; Lantolf, 2000 & 2009; Johnson, 2006 & 2009; Richards, 2008). Meanwhile, the cognitive shift in the study of teachers' language competence has changed the perspective of concept definition. Language teaching is no longer simply processes in which teachers use professional knowledge and skills, but also a complex cognitive-driven process. Teachers' language awareness, thinking activities, and creative ideas will influence this process.

In this period, performance assessment was applied in language teachers' evaluation for its advantage in observing and measuring teaching practice (Wise, Darling-Hammond & Purnell, 1988; Haertel, 1990; Scriven, 1996; Shulman, 1987; Stodolsky, 1990). Performance assessment integrates language teachers' certification with their practice "in a way consistent with both current theories of instruction and goals for education" (Arter & Spandel, 1992, p.36). Efforts regarding teachers' cognitive structures and learning processes are increasing, consistent with the increasing accentuation of performance assessment in language teaching.

2.4 A Review of Domestic Studies

With *English* as a main subject in China's school curriculum, English teachers' qualification is of great significance for the quality of English education.

Since 1990s, China's foreign language education reform triggered English teachers' professionalization and professionalism, aiming at broadening their horizons and enlarging their knowledge base. It was pointed out that "foreign language teachers should have sufficient knowledge of foreign language, while knowledge of Chinese

history and culture are also indispensable" (Wang, 1998, p.1; Zhou, 1998, p.38). Based on the above views of foreign language teachers' knowledge base, researchers extracted frameworks of foreign language teachers' competence by analyzing teachers' knowledge, skills, and dispositions (Han & Li, 1997; He & Kong, 1999; Xiao, 1999; Meng & Xu, 2005). The new century witnessed that language skills played a central role in foreign language teachers' professional competence. Empirical studies sprang up around 2005 in the field of China's EFL teacher education. During this period, the research scope in this field has been expanding, involving EFL teachers' decision-making, self-reflection, cognitive process, professional development and so on (Zhang, 2005; Gu, 2008; Han & Wang, 2008; Zhou, Cao & Wang, 2008). Researches on KAL expand EFL teachers' knowledge base. Meanwhile, researches on TLA explore "the relationship between EFL teachers' language awareness, teaching performance and learning effects" (Cheng, 2009; Gong, 2011; Lu, 2012; Huang, Zhao & Chen, 2016).

Over the past 20 years, the domestic researches on EFL teachers' competence have mainly been seen in the study of teacher education and development. Relevant researches pay more attention to teachers' knowledge, skills and dispositions. The more empirical researches are employed; the study in this field focuses more on teaching context and is more based on local teaching practice. However, up to now, China has not promulgated professional standards for foreign language teachers, nor has there been an examination for EFL teachers' certification.

3. Methodology

3.1 Research Design

This was a two-phase study comprising quantitative data of test scores and qualitative data of semi-structured and focus group interviews, integrated in a mixed methods design. Many-facet Rasch model made it possible to include comprehensive test variables relating to candidates, items, and raters. In the first phase, estimates were obtained for rater severity and task difficulty, and their influence was factored into the estimation of all participants' ability. All these estimates were shown on a single measurement scale as logits, which illustrated the relative status of elements within one facet, and the relationships between facets. The analysis of quantitative data collected in the first phase yielded typical cases for the second phase. Follow-up qualitive interviews with these cases provided insight about the performance of candidates and raters, rating scale items function, and views of stakeholders about China's EFL education, etc.

The research design provides a linking of the data sets from the two phases, during which different specific warrants are articulated and backing is amply collected. Such an approach serves as a process of assessment justification which provides a basis for accountability.

3.2 The Instrument

The ELTCertificate was introduced in 2015 at Beijing Foreign Studies University by CLA, a language assessment service integrating education and exams, and expertise and research. To meet domestic needs, the certificate was developed as a standardized assessment of ELT competencies for pre-service and in-service EFL teachers. The executed performance assessment constitutes a way of realizing a close connection between a test situation and actual classroom teaching practices so that inferences could be drawn from the test data for qualifying for ELT positions.

The frame of reference for the ELTCertificate construct definition are needs analysis of ELT tasks, theories of ELT competence, and relevant standards and syllabus (Bachman & Palmer, 2010). Based on the frame of reference, the construct of ELTCertificate is defined as English-for-Teaching and EFL Teaching competencies. Two specialized subsets of knowledge and skills are identified to distinguish particular components of the core competencies to be assessed. One is language awareness and English language skills; the other is pedagogical content knowledge and EFL teaching skills. The core competencies are found to include facilitating teacher-student interaction, assessing students' performance and giving feedback, text analysis, and lesson observation and reflection, etc. The above insights bring new perspectives on the assessment, via which performance tasks are suggested for differentiating teacher candidates' proficiency.

In the test design, Lesson Observation and Report Task (ELTC-LORT) is believed to provide information for EFL teachers' ability of oral report and evaluation on classroom ELT practice. ELTC-LORT consists of a video excerpt from a school English lesson, lasting for around ten minutes. After watching the excerpt closely, candidates have five minutes to prepare and three minutes to report. The report should include information on description, analyzing and evaluation of the teacher's performance in the video excerpt. Scales and descriptors are used to judge performance.

ELTC-LORT is an integral component of the performance assessment; it accentuates the candidates can execute

in completing the tasks. The basic design criteria of performance assessment are meaningful, authentic, and valid, according to which "validity issues must be considered part of the design template when developing an authentic assessment instrument" (Wiggins, 1992). Addressing concerns pertaining to validity entails the selection of an assessment design validation process.

3.3 Data Collection

The data employed in the executed study were derived from a July 2017 trial that examined materials used for the first administration of a test, which occurred in November 2017. The candidates (N = 40) were students enrolled in their fourth year in ELT program at a national key university at the end of the academic year 2016-2017. The reason for selecting the sample was that the candidates had completed their internships in ELT or took courses contributing to the content of abilities within the authentically assessed domain. The candidates' participation in the executed study was voluntary, and the school committee approved the research. Score reports were offered to both the candidates and their school for reference with an attempt to ensure the candidates' motivation of participation.

The candidates were provided with a ten-minute video excerpt from a real EFL lesson, and required to deliver a three-minute report on the teacher's performance. Two raters graded the records of the candidates' reports using scales of the rating rubric developed to measure EFL teacher candidates' oral English proficiency (OEP for short) and ability to analyze and evaluate teaching (AET for short). The former included the dimensions of pronunciation and intonation, fluency, language use, coherence, and logic. While the latter included the dimensions of teaching objective identifying, teaching activities description, teaching activity-objective consistency analyzing, and teaching performance evaluation. (See Appendix A & B for examples of rating scales used). The assessments were made on a four-point scale for the majority of rating scale items, with exceptions of binary ones for Logic, Objective Awareness, and Report of the Relationship between Activities and Objectives. In all, for each candidate, each of the raters made eight ratings plus one overall assessment. In the executed MFRM analysis, the ratings were considered test items. One data point was from the records of Candidate 31, whose episode broke off at the 1'32" mark.

Both raters were qualified EFL teachers experienced in ELT recruitment. One is experienced in instructing elite program in a provincial key high school. The other is a college English teacher conducting research into language assessment and language education. The raters had been trained in preparation of this administration. The rating scales and descriptors were discussed under the guidance of the research project leader. Independent ratings of sample episodes were made and discussed for the raters to meet standards of consistency and interrater agreement.

3.4 Data Analysis

The Rasch-based statistic and displays for evaluating the quality of ELTC-LORT were calculated using the Minifac (Facets Student/Evaluation) 3.80.0 program (Linacre, 2017). Specifically, a Rating Scale (RS) formulation of the MFRM model was used to explore "the psychometric quality" of the ratings examined in this study (Wright & Masters, 1982; Linacre, 1989/1994). The RS formulation of the MFRM model used in this study included facets for candidate performance, rater severity, and item difficulty. The model was specified as follows:

$$Log\ (P_{nijk}/P_{nij(k-1)})=B_n-D_i-C_j-F_k$$

Where Log $[P_{nijk}/P_{nij(k-1)}]$=the probability that performance *n* on item *i* rated by rater *j* receives a rating in catergory *k* rather than category *k-1*;

B_n=the logit-scale location (achievement) of performance *n*;

D_i=the logit-scale location (difficulty) of item *i*;

C_j=the logit-scale location (severity) of rater *j*;

F_k=the location on the logit scale where rating scale categories *k* and *k-1* are equally probable.

The category of indices and displays based on MFRM is logit scale locations. In the context of ELTC-LORT, these indices provide a method for summarizing candidate performance, rater severity, and item difficulty on a single linear scale that represents the latent construct.

The specified facets of interest in this study were candidate performance, rater severity and item difficulty. Given the multidimensionality of the task, data were entered into two data files (one OEP format file and one AET format file) with Microsoft Excel. Rasch Model statistics were calculated by inputting the files into Minifac. The output generated three categories of indices and displays which included logit scale locations, separation and

model data fit. The output of the MFRM analysis reports measurements of candidate performance, rater severity, item difficulty and bias analysis for rater×candidate interactions, rater×item interactions and candidate×item interactions.

A following-up interview is conducted aiming to triangulate the MFRM analysis. This study is framed as primarily abductive when considering interview analysis at an overarching level. Abduction is one form of reasoning employed in situations of uncertainty when we need an understanding or explanation of something that is initially diffuse (Brinkmann & Kvale, 2018, p.119). When we analyze the data collected via the MFRM and initiate to explore their further meaning, this approach is adopted to identify underlying causes.

4. Results

4.1 The MFRM Analysis of OEP

The MFRM analysis presented a measurable data summary of ELTC-LORT (OEP) candidates (n=39), raters (n=2), and rating scale dimensions (n=5). The measurable data summary demonstrated a successful estimation, with Resd (.00) and StRes (.00) as expected and 79.88% variance explained by the measures.

The vertical rulers illustrated in Figure 1 graphically displayed the latent variables investigated in this study. Specifically, it contained the calibrations of facets included in the model on the same linear scale. The labels at the top included the facets and directionality of the measure. Column 1 indicated the units of the logit scale whereby all facets can be calibrated and compared.

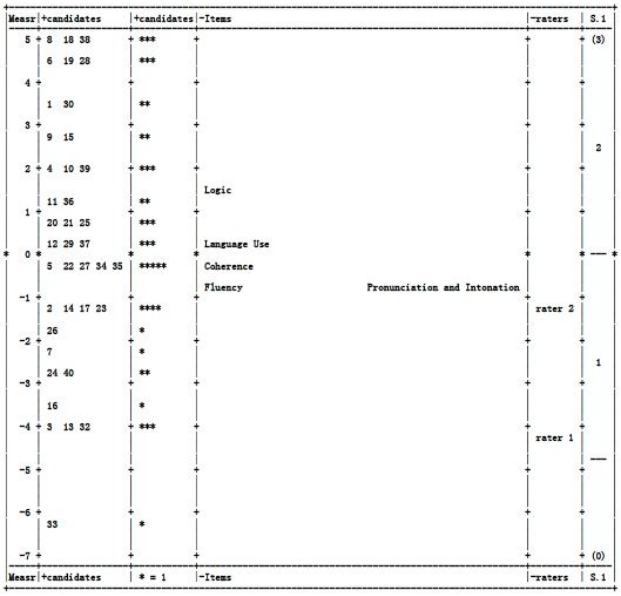

Figure 1. All facet vertical rulers of ELTC-LORT (OEP)

a Candidates: Column 2 & 3 in Figure 1 included the spread of candidates measures (i.e. performance achievement), where each asterisk represented one candidate performance. Performance achievement ranged from −6.15 to +5.89 logits (M=0.36, SD=0.44, N=39). Higher measures indicate higher performance achievement.

b Items: Column 4 included the calibrations used to represent the performance of the rating-scale, which ranged from −1.61 logits to +.82 logits. The average measures advanced largely in line with Rasch expectations. This indicated the rating scale functioned well.

c Raters: Column 5 represented the calibration of the rater measures (i.e. rater severity). Rater severity ranged from – 4.16 logits (Rater 1 more lenient) to – 1.27 logits (Rater 2 more severe).

d Bias analyses: A FACETS bias analysis allows us to "identify patterns that suggest a consistent deviation from what we would expect given the total response matrix" (Lynch and McNamara, 1998). In this study, subpatterns of bias were determined pertaining to pairs of raters with candidates, candidates with items, and raters with items.

Rater×candidate bias

The rater-candidate interactions identify questions such as particular raters' overgenerous or harsh treatment for

some candidates. Zscore controls which elements are listed. In the MFRM analysis, biases greater in size than 1 logit or with $|z|>2$ are held to demonstrate significant bias, and the null hypothesis of "the same leniency" is not rejected until $p<.05$. Since double ratings were done in this study, a pairwise report was produced to contrast the pair of raters' local behavior in the biased ratings. Table 1 reported significantly biased rater-candidate interactions.

Table 1. Bias analysis: RATER×CANDIDATE interactions (OEP)

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	+Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	raters Sq	N raters	−measr	candidates Nu can	+measr
3	6.65	5	−.73	−4.25	1.07	−3.96	4	.0167	.8	.6	26	2 rater 2	−1.27	14 lys	−1.27
7	4.13	5	.57	3.00	.96	3.13	4	.0353	.7	.6	58	2 rater 2	−1.27	32 xlm	−3.95
9.1	9.09	5.0	.00	−.10	1.20	−.13			.5	.4	Mean (Count: 72)				
2.9	2.60	.0	.25	1.48	.38	1.30			.5	.5	S.D. (Population)				
2.9	2.62	.0	.25	1.49	.39	1.30			.5	.5	S.D. (Sample)				

Fixed (all = 0) chi-square: 122.0 d.f.: 72 significance (probability): .00

* Significantly biased interactions are given.

Table 1 presents the results derived for the rater×candidate bias/interaction. It compares the raters' local behavior with their general behavior for the entirety of the dataset. Because the table presents only the most noticeable interactions, it shows that Rater 2 and Candidates 14 have an emphatic interaction, as do Rater 2 and Candidate 32. The first entry in Table 1 is for Rater 2 and Candidate 14. The observed ratings are -.73 rating points lower than we expected, which means Rater 2 has -.73 rating-points local severity. The bias size presents that this severity is -4.25 logits, with precision 1.07 logits. A test of the hypothesis has a t=-3.96, so that p=.0167 (two-sided). The null hypothesis regarding "the same severity" is rejected because Rater 2 is locally more severe than usual, which is statistically significant. The second entry is for Rater 2 and Candidate 32, which shows that Rater 2 is locally more lenient to Candidate 32 than usual by 3.00 logits, which is also statistically significant with p=.0353 (two-sided). Table 1 suggests that one rater was involved in two bias interactions, which implies that range-finding meetings and training programs should be required to ensure raters' quality.

Table 2 contrasts local behavior of the pair of raters in the biased ratings. In the context of Candidate 14, Rater 1 is 8.10 logits more lenient than Rater 2, with a paired t-test showing that the pair of raters' change in leniency is highly significant, p=.0044. Rater 1 retains conspicuous leniency when the target is Candidate 26, with 4.58 logits target contrast against Rater 2 (p=.0229). The biases shown in the second half of Table 2 report that Rater 1 is more severe than Rater 2 with Candidate 20 (-4.97 logits, p=.0119) and Candidate 32 (-6.09 logits, p=.0059). Further research is necessary to explore the pattern of the inconsistent interactions.

Table 2. Bias analysis: Pairwise report of RATER×CANDIDATE interactions (OEP)

Target Nu can	+Target Measr	S.E.	Obs-Exp Average	Context N raters	+Target Measr	S.E.	Obs-Exp Average	Context N raters	+Target Joint Contrast	S.E.	Rasch-Welch t	d.f.	Prob.
14 lys	2.58	1.64	.73	1 rater 1	−5.51	1.07	−.73	2 rater 2	8.10	1.96	4.13	7	.0044
26 wxy	.43	.98	.42	1 rater 1	−4.15	1.23	−.42	2 rater 2	4.58	1.58	2.90	7	.0229
20 sxn	−1.66	1.10	−.47	1 rater 1	3.31	.98	.48	2 rater 2	−4.97	1.47	−3.37	7	.0119
32 xlm	−7.04	1.23	−.57	1 rater 1	−.95	.96	.57	2 rater 2	−6.09	1.56	−3.89	7	.0059

* Significantly biased interactions are given.

Rater×item bias

One more model specification is put into our specification file for rater×item bias/interaction analysis. Table 3 reports the estimation of the bias/interaction, where a blank line indicates no significantly biased interactions. On average, the observed ratings in the sample exceed the expected ratings by only .09 rating-points. The interactions have a Chi-square of 13.2 with significance (probability) of .21. Accordingly, these interactions probably happened by chance.

Table 3. Bias analysis: RATER×ITEM interactions (OEP)

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	+Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	raters Sq	N raters	−measr	Items N Items	−measr
65.5	65.43	36.0	.00	.03	.42	.04			.9	.9	Mean (Count: 10)				
23.6	23.40	.0	.08	.47	.07	1.15			.2	.5	S.D. (Population)				
24.8	24.67	.0	.09	.50	.07	1.21			.2	.5	S.D. (Sample)				

Fixed (all = 0) chi-square: 13.2 d.f.: 10 significance (probability): .21

Table 4 is the bias/interaction pairwise report, which displays the contrasts of the pair of raters' local behavior in the rating scale items of Language Use and Coherence. Rater 1 is 1.28 (p=.0210) logits more lenient than Rater 2 when rating the item of Language Use, while -1.29 (p=.0228) more severe for Coherence. Therefore, the local behavior by Rater 1 is more lenient than Rater 2 in rating Language Use, while the opposite is also true in their rating of Coherence. The pairwise divergence deserves further exploration.

Table 4. Bias analysis: Pairwise report of RATER×ITEM interactions (OEP)

Target N Items	-Target Measr	S.E.	Obs-Exp Average	Context N raters	-Target Measr	S.E.	Obs-Exp Average	Context N raters	-Target Joint Contrast	S.E.	t	d.f.	Rasch-Welch Prob.
3 Language use	.81	.39	-.12	1 rater 1	-.47	.38	.12	2 rater 2	1.28	.54	2.36	69	.0210
4 Coherence	-.93	.41	.12	1 rater 1	.37	.37	-.12	2 rater 2	-1.29	.56	-2.33	69	.0228

* Significantly biased interactions are given.

Candidate×item bias

Table 5 displays the values of the candidate×item bias/interaction. The existence of a bias between items and groups of persons is equivalent to investigating Differential Item Functioning (DIF). Although in this study items of Language Use and Coherence have noticeable interactions with some candidates (biases size greater than 1 logit), none of them are of statistical significance with the precise value of probability >.05.

Table 5. Bias analysis: CANDIDATE×ITEM interactions (OEP)

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	+Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	candidates Sq	Nu	can	+measr	Items N	Items	-measr
5	3.94	2	.53	2.70	1.72	1.57	1	.3617	4.0	4.1	103	34	ys	-.27	3	Language use	.16
5	4.16	2	.42	2.19	1.72	1.27	1	.4241	4.0	4.1	99	29	wx	.24	3	Language use	.16
3	2.47	2	.26	1.33	1.64	.82	1	.5643	3.8	3.8	101	32	xlm	-3.95	3	Language use	.16
5	4.53	2	.24	1.30	1.72	.75	1	.5899	4.0	4.1	126	20	sxn	.77	4	Coherence	-.21
4	3.47	2	.27	1.27	1.53	.83	1	.5591	2.7	2.7	132	26	wxy	-1.77	4	Coherence	-.21
4	3.52	2	.24	1.13	1.53	.74	1	.5948	2.7	2.7	85	14	lys	-1.27	3	Language use	.16
4	4.56	2	-.28	-1.35	1.53	-.88	1	.5415	2.7	2.7	11	12	lxl	.24	1	Pronunciation and intonation	-.82
2	2.63	2	-.31	-1.44	1.49	-.96	1	.5121	2.4	2.4	120	13	lyy	-3.95	4	Coherence	-.21
3	3.67	2	-.34	-1.71	1.64	-1.05	1	.4859	7.2	7.2	121	14	lys	-1.27	4	Coherence	-.21
5	5.64	2	-.32	-2.05	1.73	-1.19	1	.4461	4.0	4.1	89	19	sk	4.49	3	Language use	.16
5	5.64	2	-.32	-2.05	1.73	-1.19	1	.4461	4.0	4.1	98	28	wx	4.49	3	Language use	.16
3.6	3.64	2.0	.00	-.09	1.92	-.06			.6	.6	Mean (Count: 180)						
1.6	1.51	.0	.22	1.09	.95	.65			1.1	1.1	S.D. (Population)						
1.6	1.52	.0	.22	1.09	.95	.66			1.1	1.1	S.D. (Sample)						

Fixed (all = 0) chi-square: 77.6 d.f.: 180 significance (probability): 1.00

* Significantly biased interactions are given.

The values of statistics in Items Measurement Report cause us real concern. The Outfit MnSq for the item of Language Use is 2.33, which is large enough to be distorting with Zstd significance of 4.2. Figure 2 picturesquely demonstrates the loud "noise" that drowns the "music" out. The accuracy of this item may be doubted and further checked.

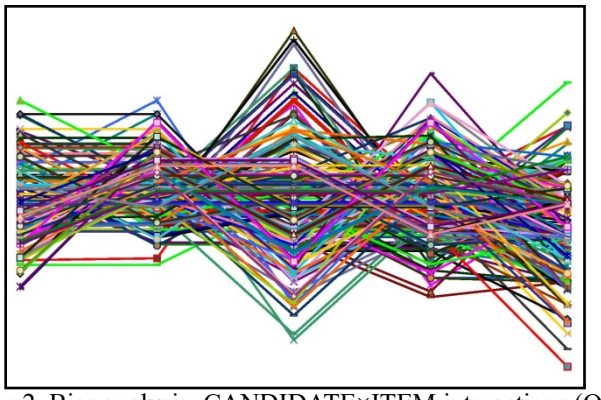

Figure 2. Bias analysis: CANDIDATE×ITEM interactions (OEP)

4.2 The MFRM Analysis of AET

The following All Facet Vertical Rulers displays a measurable data summary of ELTC-LORT (AET) candidates (n=39), raters (n=2), and rating scale dimensions (n=3). The measurable data summary presents Resd (.00), StRes (-.02), and 80.05% variance explained by Rasch measures, which gives an index to a promising estimation.

Column 6 represents the rating scale structure. Inspection the rating scale structure based upon the quantitative guidelines set forth by Linacre (1999, 2002) indicates cooperation of the rating scale categories to produce

meaningful measures.

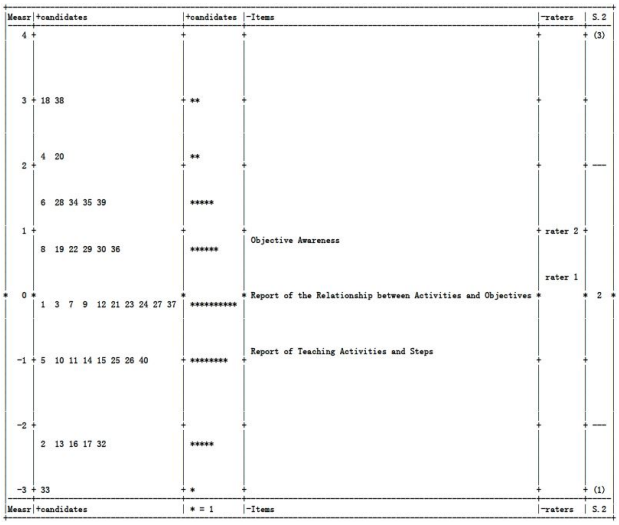

Figure 3. All facet vertical rulers of ELTC-LORT (AET)

a Candidates: There existed a noticeable spread among the candidates (–3.94 to +3.03 logits) and their traits, proving the task functioned well in measuring candidates' performance and reflecting the difference of their ability.

b Items: The rating scales worked admirably, with average measures advancing in line with Rasch expectations (–.81 to +.86 logits). The probability curves triangulated the result.

c Raters: The rater measures of Column 5 shows that Rater 2 (1.06 logits) keeps her severity as in the scoring of OEP, bearing .75 logits more than Rater 1 (.31 logits).

d Bias analyses: The iteration report ends with the Bias/Interaction analysis. The analytical model estimates the bias interactions between facets of rater, candidate, and item. The following tables display interactions with bias size bigger than 1 logit or with $|z|>2$.

Rater×candidate bias

Table 6 demonstrates significant rater-candidate interactions, but with no entry bearing a probability of $p<.05$, the null hypothesis of "the same leniency" cannot be rejected. The hypothesis was subjected to a chi-square test, and the corresponding probability was 1.00. Consequently, the hypothesis of no bias overall cannot be rejected.

Table 6. Bias analysis: RATER×CANDIDATE interactions (AET)

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	+Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	raters Sq	N	raters	candidates −measr	Nu	can	+measr
4	2.74	3	.42	1.76	1.24	1.42	2	.2909	1.4	1.7	38	2	rater 2	1.06	19	sk	.67
4	3.22	3	.26	1.07	1.24	.87	2	.4773	1.4	1.7	56	2	rater 2	1.06	28	wx	1.36
2	2.73	3	-.24	-1.34	1.50	-.89	2	.4659	3.0	2.8	13	1	rater 1	.31	7	dxc	-.09
2.7	2.66	3.0	.00	-.01	1.43	.03			.7	.6	Mean (Count: 76)						
1.1	.84	.0	.22	1.05	.24	.75			.7	.7	S.D. (Population)						
1.1	.84	.0	.22	1.06	.24	.76			.7	.7	S.D. (Sample)						

Fixed (all = 0) chi-square: 43.1 d.f.: 76 significance (probability): 1.00

* Significantly biased interactions are given.

Rater×item bias

Table 7 demonstrates a Chi-square of 34.8 with a Bias/Interaction Significance of .00, according to which the study won't allege that the interactions just happened by chance, without examining the pattern of small interactions of significance ($p<.05$). All the following four entries display significant bias with values of probability <.05 and bias size >1. Both raters' local behavior contrasts on their general behavior on the items of Report of the Relationship between Activities and Objectives and Report of Teaching Activities and Steps.

Table 7. Bias analysis: RATER×ITEM interactions (AET)

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	+Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	raters SqN	raters	-measr	N	Dimensions	-measr
24	16.47	38	.20	1.11	.39	2.83	37	.0075	.9	.8	3	1 rater 1	.31	2	Report of the Relationship between Activities and Objectives	.05
83	74.03	38	.24	.95	.33	2.91	37	.0060	1.1	1.1	6	2 rater 2	1.06	3	Report of Teaching Activities and Steps	-.86
72	81.07	38	-.24	-.96	.33	-2.95	37	.0055	.7	.7	5	1 rater 1	.31	3	Report of Teaching Activities and Steps	-.86
4	11.68	38	-.20	-1.72	.59	-2.95	37	.0056	1.0	.8	4	2 rater 2	1.06	2	Report of the Relationship between Activities and Objectives	.05
33.7	33.74	38.0	.00	-.13	.42	-.05			.9	.8	Mean (Count: 6)					
31.8	31.16	.0	.18	1.01	.09	2.41			.1	.2	S.D. (Population)					
34.8	34.13	.0	.20	1.11	.10	2.64			.1	.2	S.D. (Sample)					

Fixed (all = 0) chi-square: 34.8 d.f.: 6 significance (probability): .00

* Significantly biased interactions are given.

On the item of Report of Teaching Activities and Steps, Rater 1 (-.11 logits) performs 1.92 logits (p=.0001) more leniency than Rater 2 (1.81 logits). While on the item of Report of the Relationship between Activities and Objectives, Rater 1 (1.07 logits) performs –2.84 logits (*p*=.0001) more severity than Rater 2 (-1.77). The results lead the study to explore further why raters varying in EFL teaching background place various emphases on particular rating items.

Table 8. Bias analysis: Pairwise report of RATER×ITEM interactions (AET)

Target N	Dimensions	-Target Measr	S.E.	Obs-Exp Average	Context N	raters	-Target Measr	S.E.	Obs-Exp Average	Context N	raters	-Target Contrast	Joint S.E.	Rasch-Welch t	d.f.	Prob.
3	Report of Teaching Activities and Steps	.11	.33	-.24	1	rater 1	-1.81	.33	.24	2	rater 2	1.92	.46	4.15	73	.0001
2	Report of the Relationship between Activities and Objectives	-1.07	.39	.20	1	rater 1	1.77	.59	-.20	2	rater 2	-2.84	.71	-4.02	71	.0001

* Significantly biased interactions are given.

Candidate×item bias

In the report of candidate×item bias/interaction, all eight entries demonstrate the interactions between candidates and the very item of Report of Teaching Activities and Steps. The bias size is large enough to be noticeable (>1), while the MnSq size is not large enough to be distorting (<1.5) with probability >.05. The mean-square fit statistics show that this item is still productive. The chi-square test of hypothesis further justifies this argument with the probability of 1.00. Table 9 would suggest that the candidates were confused about teaching 'activities and steps' criterion, which may lie in their paucity of teaching practice and unfamiliarity with school syllabus.

Table 9. Bias analysis: CANDIDATE×ITEM interactions (AET)

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	+Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	candidates Sq	Nu	can	+measr	N	Dimensions	-measr
5	4.04	2	.48	1.87	1.38	1.36	1	.4035	1.4	1.4	77	1	bxl	-.09	3	Report of Teaching Activities and Steps	-.86
5	4.04	2	.48	1.87	1.38	1.36	1	.4035	1.4	1.4	111	37	zy	-.09	3	Report of Teaching Activities and Steps	-.86
5	4.41	2	.29	1.11	1.38	.81	1	.5679	1.4	1.4	84	8	hyz	.67	3	Report of Teaching Activities and Steps	-.86
5	4.41	2	.29	1.11	1.38	.81	1	.5679	1.4	1.4	95	19	sk	.67	3	Report of Teaching Activities and Steps	-.86
5	4.41	2	.29	1.11	1.38	.81	1	.5679	1.4	1.4	106	30	wyn	.67	3	Report of Teaching Activities and Steps	-.86
3	4.04	2	-.52	-2.03	1.38	-1.48	1	.3789	1.4	1.4	79	3	cyy	-.09	3	Report of Teaching Activities and Steps	-.86
3	4.04	2	-.52	-2.03	1.38	-1.48	1	.3789	1.4	1.4	83	7	dxc	-.09	3	Report of Teaching Activities and Steps	-.86
3	4.04	2	-.52	-2.03	1.38	-1.48	1	.3789	1.4	1.4	85	9	hqy	-.09	3	Report of Teaching Activities and Steps	-.86
1.8	1.78	2.0	.00	.10	1.84	.07			.6	.6	Mean (Count: 114)						
1.8	1.73	.0	.24	.91	.79	.62			.6	.6	S.D. (Population)						
1.8	1.73	.0	.24	.91	.80	.63			.6	.6	S.D. (Sample)						

Fixed (all = 0) chi-square: 45.1 d.f.: 114 significance (probability): 1.00

* Significantly biased interactions are given.

4.3 The Interview Analysis

The MFRM analysis indicates that Rater 2 acts more severely than Rater 1 both in OEP and AET, with statistically significant local severity on the dimensions of Language Use and Report of Teaching Activities and Steps. The logit-scale locations of Rater 1 rest in lower condition on both vertical rulers displaying leniency, but her local behavior in Coherence and Report of the Relationship between Activities and Objectives is more severe than Rater 2.

Semi-structured interviews and focus group interviews are employed to provide knowledge for the raters' performance and rating scale items function. The semi-structured interview was intended to obtain the meaning of rating scale items interpreted by Rater 1 and Rater 2 with respect to revealing their local behavior. In the focus group discussions, the participants, with rich experience in EFL teaching or teacher education, shared their visions on the construct specified in the rating scale items, and embodying the core competencies for new EFL teachers in China's primary and secondary schools.

Focus group interviews were conducted three times, with 19 participants (Note 3), including EFL teacher educators, EFL education researchers, primary and secondary school EFL teachers, and supervisors of the Good Teacher Project in the foreign language education field. The following results present views of the aforementioned stakeholders of China's EFL education, with the subjects remaining anonymous to protect their

confidentiality.

As a high school senior EFL teacher, Rater 2 believes that qualified EFL teachers should possess knowledge about both language and EFL teaching, bear personal teaching beliefs, disposition, as well as lifelong learning motivation. She emphasized the importance of praxis and pointed out:

> "*The most important thing for a pre-service EFL teacher is to 'follow a steady course', that is, to have qualified teaching ability*. *Knowledge about language (KAL) is continuously accumulated and gradually internalized, including structural knowledge and applied knowledge.*"
>
> (Rater 2, Semi-Structured Interview)

She outlined structural knowledge and applied knowledge in a dynamic perspective. The former involves phonetics, vocabulary, grammar, discourse (organization and genre), etc.; the latter includes knowledge of discourse construction, pragmatics, and strategies, etc.

Besides college English teacher, Rater 1 is a researcher in the field of language assessment and language education. She stressed that language proficiency tested by the Certificate should reflect levels of English-for-Teaching abilities. The construct of the certification is professional competence of EFL teaching, which is not expected to be fostered easily by candidates in their career. She further interpreted the professional competence of English-for-Teaching:

> *"This ability reflects features of EFL educators, and entails Pedagogical Content Knowledge (PCK), integrating subject knowledge and teaching knowledge, and successfully applied in EFL teaching practice."*
>
> (Rater 1, Semi-Structured Interview)

The findings of MFRM show that the rating scale item of Language Use has noticeable interactions with some candidates. Meanwhile, the bias size of Report of Teaching Activities and Steps is also large enough to be noticeable. The function of both items requires further inquiry. Focus group interviews were employed to probe into these two subsets of knowledge and skills of the core competencies constructed for the Certificate.

Participants in the focus group interviews agreed that EFL teachers must be qualified in language proficiency, and more importantly, bear the ability of language use in teaching practice, that is, the ability to use English to complete routine teaching. In the first focus group interview, Participant FB mentioned being a member of the expert panel for *the National Senior High School English Curriculum (New Version)*. After briefly interpreting the standards of language proficiency in the Curriculum, she further explained her views on the ability of language teaching:

> *"The core competencies of EFL teachers should rest with the ability of language teaching, essentials of which are filtered and refined by China's EFL education goals. This ability is reflected in language awareness, language use, and language reciprocity."*
>
> (Participant FB, Focus Group Interview 1)

In the third focus group interview, Participant FO, an academic researcher of EFL teacher cognition and development, held the opinion that:

> *"EFL teaching has its own disciplinary nature, which requires that teachers fully understand teaching objectives, grasp PCK, and apply teaching methodologies and strategies. In particular, they must know well their students and teaching environment with autonomy and reflection."*
>
> (Participant FO, Focus Group Interview 3)

The results of the qualitative interviews with the raters and stakeholders verify the construct of the Certificate specified in the rating scale items. In semi-structured interviews and focus group discussions, the participants shared their visions on the core competencies for new EFL teachers in China's primary and secondary schools. The data analysis gave rise to common themes, namely English-for-Teaching and EFL Teaching competencies. Specialized subsets of knowledge and skills are generalized, among which language awareness and English language skills for the former, and PCK and EFL teaching skills for the latter. The rating scale items were identified for the enacting of the core competencies. The items of Language Use and Report of Teaching Activities and Steps are justified to reflect KAL and PCK that distinguished the language teaching abilities of EFL teachers.

5. Discussion

These findings not only justify the use of ELTCertificate in practice, but also trigger researchers to think about the assessment and certification of China's EFL teachers' professional competence. The MFRM analysis validates the construct validity of ELTC-LORT to a certain extent. The candidates varied in level of performance, but overall, showed qualified oral English proficiency and limited EFL teaching ability. The rating scale items did vary in difficulty but function well. It was more difficult for the raters to score the items of Language Use and Report of Teaching Activities and Steps.

The working definition of ELTCertificate construct is English-for-Teaching and EFL Teaching competencies. ELTC-LORT is designed to provide information for EFL teachers' ability of orally reporting and evaluating on classroom teaching practice. The rating scale of oral English proficiency covers basic language skills like pronunciation, intonation, fluency, and grammar, besides which rubrics of language in use and discourse organization are also measured, including wording, structure, coherence, and logic. Meanwhile, the rating scale measuring candidates' abilities of evaluating ELT practice include essential factors in teaching process proposing requirements of teaching objectives identification, teaching process depiction, and teaching activities justification.

The construct of English-for-Teaching "repositions English as a practical communicative tool to carry out certain defined responsibilities within a professional or work context, the language classroom" (Freeman *et al*., 2015, p.134). For EFL teachers, basic language skills are not equivalent to English-for-Teaching, that is, the ability to use English to complete routine teaching tasks. This is largely determined by the particularity of ELT practice, that is, language is both medium and content in the teaching process (Andrews, 2001, p.75). Therefore, EFL teachers need to grasp both comprehensive language skills and the ability to interpret language in order to develop students' abilities of language use.

In exploring the EFL teaching competencies, the epistemology of education guides the study to field of practice and nature of education. The design of ELTC-LORT is not limited to talking about the discrete teaching skills, but emphasizes the integrity of classroom teaching and the ability of practice interpreting. The above results show that there exist inherent connections between factors in the overall EFL teaching practice. On the basis of mastering PCK, pre-service teachers should first of all identify teaching objectives. Teachers' understanding of teaching objectives directly affects their abilities in the use of teaching resources, the choice of teaching strategies, and the management of teaching procedures, etc. 'Presupposed objectives' will lead English teaching into the technical teaching procedures of input, processing and output of language, which fail to "meet the cognitive and emotional needs of students" (Wajnryb, 1992). 'Generated objectives' will motivate teachers to generate new objectives constantly according to learners' needs in the student-centered classroom, making ELT a meaningful process to promote language learning.

EFL teachers should not only be able to prepare, implement and monitor teaching procedures, but also to cultivate students in the integral, dynamic and interactive teaching practice. Teachers should set teaching objectives, carry out teaching activities, guide and motivate learning, and ensure the learning effect, all based on students' needs. This kind of teaching ability is embodied in teachers' 'teaching', but oriented by students' 'learning'. The problems faced by pre-service EFL teachers in class are largely due to their insufficient teaching practice. Self-observation and peer observation in field are effective way to improve their teaching performance. To give full play of classroom observation, candidates should clarify the purpose of observation, determine the focus of exploration, evaluate the organization of class and reflect on the process of their own teaching. In this regard, teaching evaluation and reflection are key performances manifesting candidates' present teaching ability and future developing potentiality.

MFRM justifies the construct of ELTC-LORT in this study to a certain degree, but the study raises concerns about the candidates' comparatively weak performance on scales of Language Use and Objective Awareness, and the raters' inconsistent local behavior on scales of Language Use and Report of Teaching Activities and Steps. To improve the validity of this performance assessment task, it is necessary to further clarify the content of the rating scales based on professional standards. However, apart from the reasons for individual differences, one crucial cause of the above performances is the absence of the professional standards for China's EFL teachers' qualification, which will provide reference for education of pre-service teachers, norms for their certification and recruitment, and guidelines for their professional development (Chen, 2008, p.42). To meet domestic EFL education needs, departments concerned should establish an EFL teacher qualification system and encourage assessment institutions to draw up professional standards for EFL teacher certification.

The enlightenment of this study is that when formulating the professional standards of China's EFL teachers and

carrying out the ELTCertificate, it is necessary to consider the actual needs of English education in China. Upon drawing close to the candidates' ELT abilities through ELTC-LORT, the study indicates that basic requirements for qualified EFL teachers should be their abilities of language use with proficient English language skills, otherwise they won't be able to provide students with correct language input. Secondly, qualified EFL teachers should bear language awareness, lack of which will not only affect their own English language teaching, but also affect their guidance to students' language development. Thirdly, it is necessary for them to have the ability to implement, evaluate, and reflect on teaching, which constitute the core professional competence of EFL teachers and are the key content of the Certificate. How to facilitate pre-service EFL teachers to grasp PCK and improve their teaching abilities is not an issue that can be solved by certification itself, but an issue that must be paid attention to in the education and professional development of EFL teachers. Therefore, the development and implementation of the ELTCertificate will have notable effects on stakeholders and improve teachers' teaching practices. These consequences are precisely what the Certificate should produce.

6. Conclusion

This study has validated ELTCertificate-LORT through a process suited for many-facet Rasch measurement supplemented with interviews. The quantitative and qualitative data discussed above have implications for the design of the Certificate tasks and also justify the use of them. The data indicated that the performance-based Lesson Observation and Report Task is valid to assess the candidates' abilities to solve practical problems in English language teaching and to evaluate the process of their applying professional competence. MFRM provides specific information about candidates, items and raters, which is valuable for the test development process.

With diversified candidates of ELTCertificate, one limitation of the study concerns that its sampling only involves students in tertiary ELT program, thus other groups of pre-service EFL teachers and novice teachers will be included in follow-up studies. Although this study is beneficial for the validation of one task of the ELTCertificate, the process of its developing and designing is "iterative" that gradually approaches perfection in constant justification and revision with larger scale trials, more extensive stakeholders, and more kinds of research methods.

China's EFL education reform has put forward new requirements for learning, teaching and assessment, ranging from cultivating students' core literacy in language learning, promulgating professional standards for teachers' certification, selecting talents skilled in foreign languages and bearing global perspectives, etc. With EFL teachers' qualification as an important mission, our efforts in the development and design of ELTCertificate are destined to be meaningful for national English education and international communication.

Acknowledgements

A previous version of this paper was presented as an address at the Pacific Rim Objective Measurement Symposium (PROMS 2018), Fudan University, Shanghai, July 2018. Thanks go to Professor Trevor G. Bond for his technical advice.

Funding

The author received financial support for the project of Research on the Evaluation of Language Talents Cultivation for International Cultural Communication in Beijing (No. 19JDYYB003), which is funded by Beijing Planning Office of Philosophy and Social Science.

References

Altis J. E., Stern, H. H., & Strevens, P. (Eds.). (1983). *Applied linguistics and the preparation of second language teachers: Toward a rationale*. Washington D. C.: Georgetown University Press.

Andrews, S. (2001). The language awareness of the L2 teacher: Its impact upon pedagogical practice. *Language Awareness, 10*(2-3), 75-90. https://doi.org/10.1080/09658410108667027

Angrist, J. D., & Guryan, J. (2004). Teacher testing, teacher education, and teacher characteristics. *American Economic Review, 94*(2), 241-246. https://doi.org/10.1257/0002828041302172

Arter, J. A., & Spandel, V. (1992). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice, 11*(1), 36-44. https://doi.org/10.1111/j.1745-3992.1992.tb00230.x

Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.

Brinkmann, S., & Kvale, S. (2018). *Doing interviews* (Vol. 2). SAGE Publications Ltd.

https://doi.org/10.4135/9781529716665

Chen, X. M. (2008). Reflection on the function of teacher qualification system and its reconstruction. *Research in Educational Development, 15*(16), 42-45.

Cheng, X. T. (2009). *An analysis of English teachers' classroom discourse*. Shanghai: Shanghai Foreign Language Education Press.

Freeman, D., Katz, A., Garcia Gomez, P., & Burns, A. (2015). English-for-teaching: Rethinking teacher proficiency in the classroom. *ELT journal, 69*(2), 129-139. https://doi.org/10.1093/elt/ccu074

Freeman, D., Orzulak, M., & Morrissey, G. (2009). Assessment in second language teacher education. In J. C. Richards & A. Burns (Eds.), *The Cambridge guide to second language teacher education* (pp. 77-90). Cambridge: Cambridge University Press.

Gibbs, G. (1983). Changing students' approaches to study through classroom exercises. *New Directions for Adult and Continuing Education, 1983*(19), 83-96. https://doi.org/10.1002/ace.36719831910

Gong, Y. F. (2011). Among primary and secondary school English teachers: Formulation of relevant standards for measurement. *Journal of the Chinese Society of Education, 219*(7), 60-65.

Gu, P. Y. (2008). Understanding successful EFL teachers' cognition and professional growth: Case studies. *Foreign Languages Research, 109*(3), 39-45.

Haertel, E. (1990). Performance tests, simulations, and other methods. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 278-294). Newbury Park, CA: SAGE. https://doi.org/10.4135/9781412986250.n17

Han, G., & Li, Q. (1997). Foreign language teaching ability and its training: Practice and research on the reform of English teaching methodology in normal colleges. *Foreign Language World, 66*(2), 6-13.

Hang, G., & Wang, R. (2008). Understanding reflective practice in EFL pre-service teacher education. *A Journal of Educational Linguistics, 123*(3), 82-87.

He, G. K., & Kong, X. H. (1999). A Study on the evaluation criteria of middle school English teachers' professional quality. *Journal of South China Normal University (Social Science Edition), 1999*(1), 79-88.

Huang, L. Y., Zhao, J., & Chen, X. Y. (2016). The prediction value of Chinese EFL teachers' knowledge of basic language construct. *Foreign Language Teaching and Research, 48*(4), 583-593.

Johnson, K. E. (2006). The sociocultural turn and its challenges for second language teacher education. *TESOL Quarterly, 40*(1), 235-257. https://doi.org/10.2307/40264518

Johnson, K. E. (2009). *Second language teacher education: A sociocultural perspective*. New York: Routledge. https://doi.org/10.4324/9780203878033

Lantolf, J. P. (2000). Second language learning as a mediated process. *Language Teaching, 33*(2), 79-96. https://doi.org/10.1017/S0261444800015329

Lantolf, J. P. (2009). Knowledge of language in foreign language teacher education. *The Modern Language Journal, 93*(2), 270-274. https://doi.org/10.1111/j.1540-4781.2009.00860_4.x

Latham, A. S., Gitomer, D., & Ziomek, R. (1999). What the tests tell us about new teachers. *Educational Leadership, 56*(8), 23-26.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511815355

Linacre, J. M. (1989/1994). *Many-facet Rasch measurement*. Chicago: MESA Press.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2017). Minifac® (Facets Student/Evaluation 3.80.0) [Computer Software]. http://www.winsteps.com/minifac.htm. Beaverton, Oregon: Winsteps.com.

Lu, X. H. (2012). How foreign language teachers construct their language knowledge from the perspective of Vygotsky's sociocultural theory. *Global Education, 41*(7), 26-32, 90.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180. https://doi.org/10.1177/026553229801500202

Meng, Z., & Xu, W. Y. (2005). On foreign language teaching methodology research. *Foreign Language World, 110*(6), 23-29.

Pulverness, A. (2015). A brief history of Cambridge English Language Assessment Teaching Qualifications. In R. Wilson & M. Poulter (Eds.), *Assessing language teachers' professional skills and knowledge* (pp. 11-31). Cambridge: Cambridge University Press.

Richards, J. C. (2008). Second language teacher education today. *RELC Journal, 39*(2), 158-177. https://doi.org/10.1177/0033688208092182

Rutherford, D. (1987). Indicators of performance: Some practical suggestions. *Assessment and Evaluation in Higher Education, 12*(1), 46-55. https://doi.org/10.1080/0260293870120105

Scriven, M. (1996). Assessment in teacher education: Getting clear on the concept. *Teaching and Teacher Education, 12*(4), 443-450. https://doi.org/10.1016/0742-051X(96)81804-X

Shulman, L. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1-23. https://doi.org/10.17763/haer.57.1.j463w79r56455411

Shulman, L. S. (1988). A union of insufficiencies: Strategies for teacher assessment in a period of educational reform. *Educational Leadership, 46*(3), 36-41.

Stephens, M. M. (2018). *Examining physics teachers' formative assessment knowledge: A many-facet Rasch model approach* (Doctoral dissertation, University of Alabama Libraries).

Stodolsky, S. S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175-190). Newbury Park, CA: SAGE.

Wajnryb, R. (1992). *Classroom observation tasks: A resource book for language teachers and trainers*. Cambridge: Cambridge University Press.

Wang, Z. Y. (1998). Three important things for foreign language teachers. *Foreign Languages and Their Teaching, 77*(3), 1,15.

Weinberger, H., & Didham, C. K. (1987). *Helping prospective teachers sell themselves: The portfolio as a marketing strategy* (Paper presented at the annual meeting of the Association of Teacher Educators, Houston, TX).

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership, 49*(8), 26-33.

Wise, A., Darling-Hammond, L., & Purnell, S. (1988). *Impacts of teacher testing: State educational governance through standard-setting*. Santa Monica, CA: The RAND Corporation.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Xiao, L. Q. (1999). Foreign language teachers and foreign language teaching methodology. *Foreign Languages and Their Teaching, 128*(12), 20-23.

Zhang, L. (2005). A study of foreign language teachers' classroom decision-making: A case study of excellent foreign language teachers. *Foreign Language Teaching and Research, 37*(4), 265-270.

Zhou, L. X. (1998). Foreign language teachers should have a broad perspective. *Foreign Languages and Their Teaching, 75*(1), 38-39.

Zhou, Y., Cao, R. P., & Wang, W. F. (2008). Growing through teaching and interactions: A study of conditions and process of teacher's development. *Foreign Languages Research, 109*(3), 51-55.

Notes

Note 1. This term refers to EFL and ESL teachers in the present paper.

Note 2. Abbr. of English as a foreign language.

Note 3. Appendix C presents detailed information of the participants in the focus group interviews.

Appendix A

Rating scale for ELTC-LORT (OEP)

Level	Pronunciation and Intonation	Fluency	Language Use	Coherence	Logic
3	Pronunciation is accurate and native-like without obvious problems. Intonation is natural and native-like.	The response is fluent with few problems like hesitations, repetitions, etc.	Vocabulary is rich with some low frequency words used and it is used accurately. Sentence structures are rich with simple and complex sentences flexibly used. No obvious grammatical problems exist.	The response is coherent and well-arranged with appropriate use of rich cohesive devices.	
2	Overall, pronunciation is relatively accurate with some problems, which does not affect the comprehensibility. Overall, intonation is relatively natural though certain gap exists with native speakers.	The response is relatively fluent with some problems.	Vocabulary is not very rich but used accurately. Some teaching terminology is used. Although sentence structures are overall simple, some complex sentences are also used. Overall, relatively accurate use of sentence structures is demonstrated with individual grammatical errors.	The response is relatively coherent and well arranged with the use of some simple cohesive devices.	
1	Pronunciation is not accurate enough with more errors. Intonation is relatively appropriate with some problems and large gap exists with native speakers.	The response has poor fluency with more problems.	Simple vocabulary is used and it is basically correct except few errors. Mainly simple sentences are used with few complex ones and they are basically correct with few grammatical errors.	Less cohesive devices are used. Although the cohesive devices are simple, they are used appropriately.	The response is logic with the points of view supported by details.
0	For test-takers at this level, their oral English proficiency does not reach the standard to be an EFL teacher.				The response is not very logic. No further explanations are provided on some points of view.

Appendix B

Rating scale for ELTC-LORT (AET)

Level	Objective Awareness	Report of Teaching Activities and Steps	Report of the Relationship between Activities and Objectives
3		The response demonstrates relatively accurate description of the teaching steps and activities in a detailed way. Some aspects in the teaching process are analyzed and evaluated. Certain specific problems are pointed out with corresponding advices given.	
2		The response demonstrates relatively accurate description of the teaching activities and steps. For certain aspects, relatively detailed description is made. Simple analysis and evaluation are made on some aspects in classroom teaching. Some problems are pointed out with corresponding advices given.	
1	Correct or relatively correct teaching objectives are identified.	The response demonstrates no systematic description of the main teaching activities and steps. Only some details are included. Evaluation is made on certain aspects of the teaching activities with some advantages and disadvantages pointed out. Advices on these disadvantages are also involved.	The response relates the teaching activities to the objectives and makes simple evaluation on the achievement of the objectives.
0	The response demonstrates weak objective awareness with no teaching objectives identified.	The response randomly describes some details in teaching without systematic description of the main teaching activities and steps. Superficial analysis of some teaching details is made based on their personal learning experience. Individual problems are pointed out.	The response demonstrates no illustration of the relationship between the activities and the objectives.

Appendix C

Participants in the focus group interviews

Participants in Focus Group Interview 1

Participants	Gender	Title (Designation)	Field of Work (Research)
FA	Male	(Normal) University Professor/ EFL Education Researcher	EFL Education
FB	Female	University Professor/ EFL Education Researcher	EFL Education
FC	Male	High School EFL Teacher Educator	High School EFL Teacher Education
FD	Female	High School EFL Teacher	ELT/Teacher Recruitment/ Supervisor of the Good Teacher Project
FE	Female	Primary School EFL Teacher Educator	Primary School EFL Teacher Education
FF	Male	University Professor/ EFL Education Researcher	EFL Education
FG	Female	Primary School EFL Teacher	ELT/Teacher Recruitment

Participants in Focus Group Interview 2

Participants	Gender	Title (Designation)	Field of Work (Research)
FH	Female	Secondary School EFL Teacher	ELT/Teacher Recruitment
FI	Female	Primary School EFL Teacher	ELT/Teacher Recruitment
FJ	Female	Secondary School EFL Teacher Educator	Secondary School EFL Teacher Education
FK	Female	High School EFL Teacher	ELT/Teacher Recruitment
FL	Male	(Normal) University Professor/ EFL Education Researcher	EFL Teacher Education
FM	Male	(Normal) University Professor/ EFL Education Researcher	Teacher Education

Participants in Focus Group Interview 3

Participants	Gender	Title (Designation)	Field of Work (Research)
FN	Female	University Professor/ EFL Education Researcher	EFL Education
FO	Female	University Professor/ EFL Education Researcher	EFL Teacher Education
FP	Male	(Normal) University Professor/ EFL Education Researcher	EFL Teacher Education
FQ	Female	Secondary School EFL Teacher	ELT/Teacher Recruitment
FR	Male	Secondary School EFL Teacher Educator	High School EFL Teacher Education
FS	Female	High School EFL Teacher	ELT/Teacher Recruitment/ Supervisor of the Good Teacher Project