University English teacher assessment literacy: A survey-test report from China

Yueting Xu
Guangdong University of Foreign Studies &
The University of Hong Kong, China
Gavin T. L. Brown
The University of Auckland, New Zealand

Assessment literacy (AL) is central to the quality of education because competencies in assessing student learning lead to informed decisions. While the AL of university English teachers in China is particularly crucial as they teach the largest group of adult English language learners in the world, it has regrettably remained largely unexplored. The present study subjected an adapted version of the *Teacher Assessment Literacy Questionnaire* to rigorous psychometric property analyses, and used it to investigate the AL level of Chinese university English teachers (*N*=891) and the effects of their demographic characteristics on AL performance. Findings reveal a basic level of AL in certain dimensions with limited influence from demographic characteristics. Discussions are centered around validation of the AL instrument, causes for limited AL competence, and key factors that have impacted AL. This study concludes with a reflection of constructing contextually-grounded AL measures and implications for principles, policy and practice of teacher assessment education.

Keywords: assessment literacy; China; university English teachers

Introduction

Assessment literacy (AL), traditionally defined as a basic understanding of educational assessment and related skills (Stiggins, 1991), is increasingly recognized as an integral part of teacher expertise (Popham, 2009; Xu & Brown, 2016). It is generally agreed that teachers need a sound mastery of assessment principles and techniques to make sophisticated judgments about the validity of assessment practices and/or policies in specific contexts (Kane, 2006; Messick, 1989). Sufficient AL enables teachers to make accurate inferences about student learning, communicate that information to students

Email address for correspondence: traceyxu@gdufs.edu.cn

and other stakeholders, and adjust instruction accordingly, while insufficient AL leads to reduced reliability and validity, and further results in misdirected and ill-informed educational decisions. Hence teacher AL is arguably placed at the heart of the success of educational assessment and even the overall quality of education.

In the field of language testing and assessment, language assessment literacy (LAL) is used to refer to the AL required for various stakeholders, including language teachers. As 'a potentially subordinate or overlapping category' to AL (Taylor, 2013, p. 405), LAL is likely to have multiple layers and progressive stages (Pill & Harding, 2013; Taylor, 2013). In terms of stages, LAL for teachers seems to range from a basic understanding of measurement knowledge and assessment 'know-how' for classroom practice to a more advanced level of 'having the capacity to ask and answer critical questions about the purpose for assessment, about the fitness of the tool being used, about testing conditions, and about what is going to happen on the basis of the results' (Inbar-Lourie, 2008, p.389). Despite these discussions of LAL, one pertinent issue that has remained unresolved is what specific in-house expertise is included in LAL (Inbar-Lourie, 2013). Given the unspecified dimensions of basic LAL competencies, in this paper we use AL or teacher AL rather than LAL as the conceptual term.

Despite the compelling arguments for AL (Brookhart, 2011), many teachers are often involved in assessment decision-making without sufficient training in assessment (DeLuca & Bellara, 2013; Schafer & Lizzitz, 1987). While teachers may spend about a half to a third of their professional time on assessment-related activities (Stiggins, 1995), the status quo of teacher AL, however, is regrettably far from satisfactory (DeLuca & Klinger, 2010; Popham, 2009). Therefore, understanding teachers' current levels of AL mastery is a good departure point for promoting both AL research and teacher development in assessment.

Understanding the AL level of university English teachers in China is a particularly pressing task due to three factors. First, their AL is highly consequential as they teach the largest group of adult English language learners in the world. Second, they have enormous assessment responsibilities due to the co-existence of two competing assessment purposes (i.e., accountability and learning) (CMoE, 2007). Third, their current AL levels have remained underexplored. Compared to the burgeoning teacher AL research conducted in the 'Western' educational contexts (DeLuca, Chavez, & Cao, 2013; Fulcher, 2012; Plake, Impara, & Fager, 1993), similar studies in China are generally lacking. The present study addresses this gap by investigating the current AL level of university English teachers in China.

Teacher assessment literacy: mastering theoretical principles

Empirical evidence of teacher assessment literacy converges on three themes: knowledge and skills within AL, assessment education and its relationship with various mediating factors, and contextual considerations of AL (see a review in Xu & Brown, 2016). For the purpose of this paper, the focus of the review is restricted to the first and second strands because we believe that AL research needs to start with a substantial discussion of its knowledge base, a careful analysis of relevant measures, and a full understanding of factors that exert an impact on teacher AL.

Discussions of the AL knowledge base can be traced to a seminal document--the *Standards for Teacher Competence in Educational Assessment of Students* (hereafter the *Standards*) (AFT, NCME, & NEA, 1990). The *Standards* prescribe seven competency domains in which teachers should be skilled; that is,

1. choosing assessment methods appropriate to instructional decisions;
2. developing assessment methods appropriate to instructional decisions;
3. administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods;
4. using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement;
5. developing valid pupil grading procedures;
6. communicating assessment results to various stakeholders; and
7. recognizing unethical, illegal, and inappropriate assessment methods and uses of assessment information.

Recent studies have proposed updated lists of AL knowledge base (e.g., Brookhart, 2011; JCSEE, 2015; Stiggins, 2010), calling for inclusions of capabilities based upon recent developments in assessment policy and practice. For example, Brookhart (2011) noted that the *Standards* do not consider recent development of formative assessment, with one important aspect being self- and peer assessment. It is essential that assessment literate teachers are able to engage students in self- and peer assessment.

These updated lists of AL knowledge base notwithstanding, the *Standards* have remained the most popular blueprint for developing AL measures. These measures particularly appear in the way of objective tests of assessment knowledge, and investigate the extent to which teachers know about the prescribed assessment principles with identified strengths and weaknesses. The most widely used one was the *Teacher Assessment Literacy Questionnaire* (TALQ) (Plake, Impara, & Fager, 1993), later going by the name of *Classroom Assessment Literacy Inventory* (CALI) (Mertler, 2004). It consisted of 35 items, with every

five items measuring one competency area specified in the *Standards*. Each item goes with four options, with correct answer being dichotomously scored and a high score representing a high level of AL. Table 1 outlines the empirical results of AL and psychometric evidence for the instrument, where available.

Table 1. Psychometric Properties of TALQ and Related Measures

Authors (Year)	Instrument	Sample	Results	Psychometric properties
In-service teachers				
Plake, et al. (1993)	TALQ	555 elementary and secondary school teachers in U.S.	*M*=23.2/35 =66%	KR_{20}=.54
Mertler (2004)	CALI	101 secondary school teachers in U.S.	*M*=21.67/35 =62%	α=.44
Zheng (2010)	Abbreviated TALQ with 21 items	954 primary and secondary school teachers in China	*M* =9.56/21 =46%	Not reported
Alkharusi, Kazem, & Al-Musawai (2011)	TALQ	233 teachers in Oman	*M*=12.55/35 =36%	Not reported
Pre-service teachers				
Mertler (2004)	CALI	67 undergraduates in U.S.	*M*=18.96/35 =54%	α=.74
Mertler & Campbell (2005)	ALI	250 undergraduate in U.S.	*M*=23.83/35 =67%	KR_{20}=0.74; item difficulty (*M*= .68); item discrimination (*M* = .31)
Alkharusi (2011)	TALQ	259 undergraduates in Oman	*M*=20/35=57%; *SD*=8	KR_{20}=.84; item difficulty (*M*=.56); item discrimination (*M* = .51); CFA fit indices (χ^2 = 990.762; *df*=329; CFI=.89; RMSEA=.08)
Alkharusi, et al. (2011)	TALQ	279 undergraduates in Oman	*M*=15.30/35 =47%; *SD*=3.94	Not reported

Table 1 can be summarized as follows: 1) the absolute number of empirical studies measuring teacher AL is relatively small; 2) the average sum of items answered correctly for both pre- and in-service teachers is low to medium; 3) the reliability estimates are higher among pre-service teachers than among in-service teachers; 4) most studies were conducted in the U.S.; the only study conducted in China was administered among school teachers teaching various subjects; and 5) psychometric properties at the item level are generally poor to fair, except for Alkharusi's (2011) study which provided both classical test theory item indices and confirmatory factor analysis (CFA) of a one-factor solution and proposed that further examination of TALQ using item response theory analysis of responses with a much larger sample of participants is needed.

In addition to reports of general level, AL has often been investigated in relation to basic teacher demographics, which produced mixed results. Prior studies generally reported positive effects of assessment training on teacher AL (Alkharusi, et al., 2011; DeLuca, Chavez, & Cao, 2013; Graham, 2005; Lukin et al., 2004; Mertler, 2009). One exception to this general finding is Brown's (2008) which found that training experience in assessment had no effect on participants' AL. In addition, no clear consensus was reached on the relationship between teaching experience and AL levels, as a positive relationship was found in some studies (Hoover, 2009; Mertler, 2004) but not in others (King, 2010; Zhang & Burry-Stock, 2003). Other demographic characteristics include teacher qualification (e.g., obtaining a graduate degree) which was found to be related to a higher level of AL (Hoover, 2009; King, 2010), and schooling experience which was found to impact teachers' assessment decisions (Campbell & Evans, 2000).

Three research gaps are thus identified from the review. First, psychometric properties of items and factors in AL measures need to be examined more rigorously to properly establish a basis for any claims made based upon the measures. Second, AL measures need to be adapted to take formative assessment into considerations. Third, new evidence about the impact of teacher demographic characteristics on AL is needed due to the unresolved controversy of the issue. The present study addresses these gaps by investigating the current AL levels of university English teachers in China and influence from demographic features, as well as by subjecting the responses to advanced psychometric analyses.

Method

By adapting the TALQ, we intend to answer the following three research questions:

RQ1: What are the psychometric properties of the adapted TALQ?

RQ2: What is the general AL level of university English teachers in China?
RQ3: Do teachers' demographic characteristics (i.e., gender, age, professional title, qualification, years of teaching, assessment training, university level, and region) have an effect on their AL performance?

Research context

China is a geographically vast country, conventionally divided into seven regions (north, south, central, east, northeast, northwest, and southwest). There are three types of universities: (1) first-tier universities which are sponsored by the "985" and "211" projects, as well as a few provincial key universities, (2) second-tier universities which are non-key universities that enroll four-year undergraduates, and (3) local community colleges which are vocationally oriented and offer three-year training programs. The "211" and "985" projects are two governmental initiatives that are aimed, respectively, at strengthening about 100 tertiary institutions and establishing 39 world-class universities. There are approximately 191 first-tier, 596 second-tier, and 1741 local community colleges in China (Xie, 2014).

Given that there are 25,476,999 undergraduate and 1,847,689 graduate students in China (Xie, 2014) who need to learn English as a compulsory course, demand for university English teachers is huge. This demand has prompted many tertiary institutions to lower their standard when recruiting English teaching professionals. Compared to a relevant PhD degree as the minimum qualification for a teaching position in many subject departments, the threshold for becoming a university English lecturer in China is a relevant master's degree, although requirements vary depending on the university level, as well as the social economic status of the city in which the university is located. This is confirmed by Wang and Wang's (2011) national survey which reported that doctorate and master's degree holders respectively take up 1.5 and 60.1 per cent of the population of university English teachers in China.

At the policy level, specialized assessment policies for higher education in China do not exist. The only document that explicitly states the requirements for university English teachers' assessment practice is the College English Curriculum Requirements (hereafter the CECR) (CMoE, 2007) which prescribes the parallel positioning of summative and formative assessment. Since the promulgation of the CECR, formative assessment has been increasingly used by university English teachers, although their practices are reported to be heavily influenced by the mindset of the testing culture (Chen, May, Klenowski, & Kettle, 2014). Thus, assessment practice in university English language teaching can be described as nominally formative, but practically speaking, it is summative mimicry of the examination system.

Participants

A random stratified sampling strategy was adopted to identify target universities. Since China is a geographically vast country, it was not feasible to collect data from every province and municipality. Thus, the conventional geographical division of seven regions of China was used as one stratum. The second stratum was the ranking: first-tier universities, second-tier universities, and local community colleges.

These two strata produced 21 cells, from which one institution was chosen randomly. Invitations to participate were sent to deans or division heads of these institutions, which led to a 43% acceptance rate. When the invitation was declined or no response was received, an alternative institution from the same cell was invited, resulting in an overall 33% acceptance rate. In total, the recruitment went through nine rounds before all the 21 cells were filled.

Institutions that agreed to participate chose either to disseminate the survey by paper or online. Paper-based surveys were administered at a faculty meeting where the research project was introduced to the teachers, while online administration was conducted through email invitations for completion on a survey website (www.sojump.com). For both types of administration, participation was voluntary. The return rates for paper-based and Internet-based surveys were respectively 94% of the 210 hard copies distributed on site and 79% of the 900 invitations sent. This high return rate ensures the low potential of non-response error (Dillman, 1991). Although the vast majority of responses (80%) were obtained from the Internet administration, method of administration had little impact on response time, with an average of 25 minutes for both types of administration.

Table 2. Participant Demographic Information

Demographic Category	*N*	%
Region		
Central China	112	12.6
East China	239	26.8
North China	139	15.6
South China	275	30.9
West China	126	14.1

Demographic Category	*N*	%
Gender		
Men	199	22.3
Women	692	77.7
Age		
Below 30	73	8.2
30 to 39	501	56.2
40 to 49	263	29.5
50 to 59	54	6.1
Qualification		
Bachelor	78	8.8
Master	679	76.2
Doctorate	134	15
Title		
Teaching assistant	66	7.4
Lecturer	536	60.2
Associate professor	252	28.3
Professor	37	4.2
Years of teaching		
less than 3 years	65	7.3
4 to 6 years	93	10.4
7 to 15 years	439	49.3
15 to 25 years	216	24.2
more than 25 years	78	8.8
University		
First-tier	457	51.3
Second-tier	322	36.1
Local community college	112	12.6
Assessment training		
None	381	42.8
Pre-service Only	240	26.9
In-service Only	74	8.3
Training Pre- + In-service	196	22
Data Collection Method		
Paper	188	21.1
Internet	703	78.9

The participants were a large sample of university English teachers working across China (*N*=891). The margin of error for this sample size relative to the population of university

English teachers (*N*=130,601) (Xie, 2014) was estimated as 3.27% with a 95% level of confidence, which supports the generalizability of the sample to the population. As Table 2 shows, while these respondents varied in their demographic characteristics, those who were lecturers and holding master's degrees made up the largest portion of the sample. It again confirms the qualification bar for university English teachers introduced earlier. It should also be noted that over 40 per cent of the respondents reported that they had not received any assessment training either in pre- or in-service teacher education.

Given that the responses from the seven regions were not equal, which may threaten the validity of using region as a factor in analysis of variance, 'Northwest' and 'Southwest' were aggregated into a 'West' region, while 'Northeast' was included under 'North'. Unsurprisingly, there was considerable overlap between being older and having more teaching experience (Cramer's V = .77, well beyond chance). This means that only "years of teaching" was used as a predictor when addressing RQ 3.

Instrument

Part 1 of the adapted TALQ had seven items about participant demographic characteristics (e.g., age, gender, years of teaching, highest qualification, current title, university level, and region) and two items about their prior assessment training experiences. These factors were included mainly based upon the existing literature on mediating factors of AL that were reviewed earlier. Part 2 consists of 24 dichotomously-scored items measuring the eight competency domains of teacher AL (see Table 3 for the descriptors of each competency and sample items).

Among the 24 items, 21 came from the original TALQ. The main consideration for a shortened TALQ version was for minimizing fatigue and improving response rates in light of the fact that the original TALQ usually requires more than 40 minutes to complete. The adaptation was guided by two principles: (a) retaining the AL constructs (domains) as designed in the TALQ; and (b) ensuring that the items were relevant, meaningful, and realistic to Chinese university English teachers' lived experiences. In addition to the new construct, three major revisions to the TALQ were made:

1. To reduce reading demands, all names appearing in the scenarios were replaced with a personal hypothetical scenario starter "Suppose you...". This modification was expected to position respondents within these scenarios and to prompt them to make choices according to their own experiences.
2. The content in each scenario was changed to describe relevant materials of English language teaching and learning, while the context was changed from K-12 schooling to higher education settings. Two items specifically related to K-12

education were removed (i.e., Items 20, 33 in the original TALQ).

3. Items that had been designed specifically for the U.S. policy and practice contexts were also excluded (i.e., Items 5, 27, 31 in the original TALQ).

In addition to the seven competency domains specified in the *Standards*, we included a new competency which requires teachers to be 'skilled in using various strategies to help students become competent assessors of their own or others' work.' Each competency has three items, and altogether the adapted TALQ consists of 24 items measuring eight competency domains of assessment literacy, as specified in Table 3.

Table 3. Descriptors of the Eight Competency Domains and Sample Items in the Adapted TALQ

Competency /Item No.	Descriptors of the competency domain /Sample Item
1	Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
Item 2	*When scores from a standardized test are said to be reliable, what does it imply?*
2	Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
Item 5	*Suppose you want your students to appreciate the poems of Emily Dickinson in your course of "British and American Literature Appreciation". Which of the following test items below would best measure your instructional goal?*
3	Teachers should be skilled in administering, scoring and interpreting the results of both externally produced and teacher-produced assessment methods.
Item 7	*Suppose that students in your "Academic English" course are required to write an academic paper based on their own subject area as part of their end-of-unit grade. Which scoring procedure below will maximize the objectivity of assessing these papers?*
4	Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
Item 10	*Suppose you are starting a new semester with your second-year students in the course of "Integrated Skills of English". Before beginning the class, you give your students a test on listening comprehension, reading comprehension, grammar and vocabulary, cloze and writing. Which of the following is most likely the reason for your giving this test?*
5	Teachers should be skilled in developing valid grading procedures.
Item 15	*Which of the following grading practices leads to a grade that least reflects students' achievement?*
6	Teachers should be skilled in communicating assessment results to students, parents, other educators and general public.
Item 16	*John got 7 in all three English courses, Integrated Skills of English, English Speaking, and Writing. The scores are all Stanine scores (which is a method of scaling test scores on a nine-point standard scale with a mean of five and a standard deviation of two). Which of the following is a valid interpretation of the scores?*
7	Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.
Item 19	*Which teacher's action is considered ethical?*

Competency /Item No.	Descriptors of the competency domain /Sample Item
8	Teachers should be skilled in using various strategies to help students become competent assessors of their own or others' work.
Item 22	*Suppose that you want your students to review each other's argumentative essays. Which of the following would maximize the peer review outcomes?*

Note. *Each competency has three items.*

Following these adaptations, the first draft of the adapted TALQ was translated into Chinese. Two Ph.D. students who were native in Chinese and highly proficient in English compared the Chinese and English versions, which resulted in revisions based on their proofreading suggestions. It was then piloted sequentially with three groups of people: (a) seven Ph.D. candidates whose research areas were English language education and who had taught English at tertiary-level in China, for clarity and conciseness of wording of items; (b) three language assessment professors for content validity of items; and (c) 55 university English teachers working in a key university for appropriateness of the scenarios to their context. Based on the feedback from these three groups, it was further revised to ensure accuracy, clarity and ease of understanding before it was formally administered.

Data analyses

Data were checked for completeness before being analyzed. Ten participants were excluded for having missing values, and another ten were excluded for containing conflicting demographic information. Hence, subsequent analyses were conducted with 891 valid and complete questionnaires.

To address RQ 1, item response theory (IRT) was used to establish the validity of the test items in terms of discrimination and guessing probability. IRT is a theory of statistical estimation (de Ayala, 2009) that defines the correspondence between latent variables and their manifestation in responses to test questions. In three parameter logistic (3PL) IRT, an item's characteristics are described by an ogive-shaped curve in which the item difficulty (*b*) is determined when the curve crosses the 50% probability threshold of being answered correctly; the item discrimination (*a*) is defined as the slope of the curve at the difficulty point, and the pseudo-guessing (*c*) is determined by the probability level when θ= -3.00. Item difficulty and personal ability are mutually defined on a common scale (θ), which has a mean of zero and a standard deviation of 1.00. Under the 3PL analysis, item discrimination values should be >0.00, while item's pseudo-guessing values should be <.25, since all items had 4 options. The 3PL IRT analysis was conducted with SPSS R-Plugin.

Confirmatory factor analyses (CFA) were then conducted to determine the validity of the latent factor structure. Weighted Least Square (WLS) estimation of the variance-covariance tetrachoric correlation matrices was conducted to examine the fit of the data to the structural model (Raykov & Marcoulides, 2007) in Lavaan package of R (Rosseel, 2012). WLS rather than maximum likelihood estimation was used because of the dichotomously-scored items in the adapted TALQ (i.e., 0 or 1). Models with statistically non-significant χ^2/df, gamma hat >.90, comparative fix index (CFI) >.90, root mean square errors of approximation (RMSEA) <.08, and standardized root mean residuals (SRMR) <.06 are considered well-fitting with data (Fan & Sivo, 2007).

Once a psychometrically defensible set of test items was found, RQ 2 was answered with overall scores of the items calculated using IRT 3PL approach in PARAM (Rudner, 2012). This is superior to the classical sum of items correctly scored since it gives higher scores to persons who answer the more difficult questions correctly. With a score for each person, RQ 3 was answered with multi-factorial analysis of variance to determine whether any of the demographic variables provided a meaningful explanation for the tested AL performance.

Limitations

This study carries some limitations. First, logistics have constrained us from collecting similar numbers of questionnaires from every region of China. Thus the results might be well represented by teachers from eastern and southern parts of China, yet under-represented by those from the West and North. Second, out of concerns for cognitive loading and completion time, every construct (domain) of AL was measured by only three items. The limited number of items might lead to lower internal consistency of the items within the measure, but it was the trade-off we had to make to encourage more voluntary participation.

Results

RQ1: What are the psychometric properties of the adapted TALQ?

The overall internal consistency for the 24 items in the adapted TALQ was α=.53, consistent with previous studies conducted among in-service teachers (cf. Mertler, 2004; Plake et al., 1993). The relatively low internal consistency of these items may be due partly to the limited number of items representing each construct (domain) (i.e., three items),

and partly to the fact that in-service teachers may have greater experience that permits them to make finer distinctions among the intended constructs.

CFA was used to determine if a multi-factorial model could be fit to the data. However, the inter-correlated eight-factor CFA model was inadmissible because the covariance matrix was not positive definite; this conventionally indicates that too many factors have been specified. Inspection of the produced correlation values confirmed this analysis since five factors had correlations r>1.00 with Factor 4 (i.e., using assessment results).

A single factor model with all 24 items was tested. An admissible solution was found (χ^2 =581.918; df=276; χ^2/df=2.108, p=.15; CFI=.569; gamma hat =.97; RMSEA=.038; SRMR=.041). Table 4 provides the CFA factor loadings of each item and item properties derived from the 3PL IRT analysis. While the single-factor 24-item model had good fit to the data, eight items had very low factor loadings (i.e., λ<.20), another eight had lower than conventional loadings (i.e., .20<λ<.30). In accordance with IRT scaling of mean difficulty at 0.00, 13 items had negative theta values and 11 were positive. The difficulty range was from -2.79 to 5.20. Seven items had high pseudo-guessing values (c>.25) and two items (both in Factor 4) had negative discrimination values. Thus, nine items failed to meet IRT standards for a good item. These analyses suggested that the eight competency domains measured by 24 items of the adapted TALQ were not supported by CFA, and a trimmed model with fewer items may be warranted.

Table 4. Item Psychometric Properties

		3PL IRT characteristics			Factor Loadings
Factor	Item	Difficulty (b)	Discrimination (a)	Guessing (c)	(λ)
	q1	**-0.126**	**1.536**	**0.627**	**.26**
1	**q2**	**1.331**	**1.74**	**0.256**	**.22**
	q3	1.49	0.193	0.025	**.10**
	q4	**0.583**	**1.22**	**0.301**	**.29**
2	q5	-1.577	1.001	0.000	.39
	q6	**2.724**	**2.104**	**0.269**	**-.01**
	q7	-1.575	1.372	0.000	.44
3	q8	5.155	0.163	0.002	**.11**
	q9	-0.518	0.521	0.000	**.24**
	q10	-1.086	0.566	0.000	**.25**
4	**q11**	**-2.79**	**-0.792**	**0.226**	**-.16**
	q12	**-2.079**	**-15.529**	**0.214**	**-.08**
	q13	**1.985**	**7.988**	**0.324**	**.04**
5	q14	2.537	0.168	0.001	**.11**
	q15	-0.574	0.977	0.000	.42
6	q16	-0.298	0.665	0.000	.31

Factor	Item	3PL IRT characteristics Difficulty (b)	Discrimination (a)	Guessing (c)	Factor Loadings (λ)
	q17	**-0.788**	**1.828**	**0.434**	.41
	q18	**1.705**	**1.784**	**0.296**	**.14**
	q19	1.188	0.528	0.000	**.24**
7	q20	-0.719	0.400	0.000	**.22**
	q21	**0.131**	**0.779**	**0.477**	**.21**
	q22	0.088	0.85	0.000	.41
8	q23	-0.752	1.113	0.000	.45
	q24	-1.205	0.766	0.000	.34

Note. *Items in bold violate conventions for acceptance.*

Given that both CFA and IRT analyses indicated different items had problematic properties, multiple models with different sets of discarded items were tested. In Model 1 eight items whose factor loadings <.20 were deleted. Model 2 deleted the nine items that did not meet the IRT standards due to negative discrimination and guessing value higher than .25. Finally, Model 3 deleted all 14 items indicated as problematic by both approaches. All three models were tested in a single factor structure. Results (Table 5) show that all models had good fit to the data. Inspection of standardized differences in chi-square relative to differences in *df* (Wilson & Hilferty, 1931) indicated that Models 1 and 2 differed by no more than chance (z=1.26; p=.10), while the difference between Model 1 and Model 3 was significant (z=4.46; p<.001), indicating it was a better representation of a single latent trait underlying responses to the adapted TALQ.

Table 5. Alternative Model Fit Statistics

Model	*n*	χ^2	*df*	χ^2/df (*p*)	CFI	Gamma hat	RMSEA	SRMR
1	16	140.172	77	1.82 (.18)	.72	.99	.03	.035
2	14	177.022	104	1.70 (.19)	.76	.99	.03	.036
3	10	47.863	35	1.37 (.24)	.93	.99	.02	.031

Note. *CFI=comparative fit index; RMSEA=root mean square error of approximation; SRMR=standardised root mean residual*

RQ2: What is the general AL level of university English teachers in China?

3PL scores were created for the ten valid items. Table 6 provides the constructs, items, and psychometric details of the 10 retained items in Model 3. Factor loadings ranged between .23 and .45. Only one item (q19) was highly difficult (θ>1.00) and one (q22) was of medium difficulty, while the remaining eight were relatively easy (i.e., θ<0.00). The 10 retained items sampled seven of the eight intended content domains, five of which with only one item. Only Domain 1 (i.e., choose assessment methods appropriate for instructional decisions) had no representation, while Domain 3 (i.e., administer, score,

and interpret externally produced and teacher-produced assessment methods) had two items, and Domain 8 (i.e., involve students in assessment of their own or others' work) had three items. The easy items (θ<0.00) all suggest quite a basic level of AL competence involving aligning tasks to instructional goals, objective scoring of tasks, clarity as to the purpose of assessments, engaging students in assessment practices, valid grading, and accurate interpretation of standard scale scores. A mastery of these ten items would suggest a basic functional competence rather than advanced capability.

The mean score for the sample was -0.18 (*SD*=1.06), which indicates that, on average, participants would have a low probability (i.e., *p*<.50) of answering correctly the two difficult items, while being highly likely (i.e., *p*>.50) to answer correctly the eight easy ones. Among the participants, almost one-fifth (*n*=173, 19%) of participants got scores below (θ<-1.00), while two-fifths (*n*=374, 42%) had scores between -1.00 and 0.00. The remaining nearly two-fifths (*n*=344, 39%) had scores >0.00. This suggests that the range of competencies in AL were large, with a vast majority having very basic to minimally acceptable competencies. However, given that results of the deleted 14 items were not explored, the overall AL level of the sample of university English teachers surveyed in China cannot be determined. It can only be inferred from their performance on the 10 items that they seem to have a basic level in certain dimensions of AL, that is, aligning tasks to instructional goals, objective scoring of tasks, clarity as to the purpose of assessments, engaging students in assessment practices, valid grading, and accurate interpretation of standard scale scores.

Table 6. Retained Items[2] and Competency Domain measured ordered by Theta Value

Competency; Item(s), & Psychometric Values	Item and Key
Develop assessment methods appropriate to the instructional goal	
5. (θ=-1.61; λ=.39)	*Suppose you want your students to appreciate the poems of Emily Dickinson in your course of "British and American Literature Appreciation". Which of the following test items shown below would best measure your instructional goal?* Discuss briefly your understanding of the uniqueness of Emily Dickinson's poems.
Score results of teacher-produced assessment	
7. (θ=-1.52; λ=.44)	*Suppose that students in your "Academic English" course are required to write an academic paper based on their own subject area as part of their end-of-unit grade. Which scoring procedure below will maximize the objectivity of assessing these student papers?* Before the papers are turned in, prepare a model or blueprint of the critical features of the paper and assigns scoring weights to these features. The papers with the highest scores receive the highest grades.

[2] The complete version of the adapted TALQ is available upon request from the corresponding author.

Engage students in assessment	
24 (θ=-1.23; λ=.34)	*Which of the following may engage students in assessment?* All of the above.
Use assessment results for decision-making and planning	
10 (θ=-1.00; λ=.25)	*Suppose you are starting a new semester with your second-year students in your course entitled "Integrated Skills of English". Before beginning the class, you give your students a test on the listening comprehension, reading comprehension, grammar and vocabulary, cloze and writing. Which of the following is the most likely reason for your giving this test?* To check for prerequisite knowledge in students to inform lesson-planning and instructions before beginning the course.
Engage students in assessment	
23 (θ=-0.74; λ=.45)	*You will ask students to write seven essays as the assignments for your course English Writing. They are required to write first, second or even third draft for each essay. If you wish to ask students to use portfolio to self-assess their own writing, which of the following is least likely to help students become better self-assessors?* Ask students to decide for themselves which drafts/essays to put in the portfolio
Grading	
15 (θ=-0.52; λ=.42)	*Which of the following grading practices results in a grade that least reflects students' achievement?* To check for prerequisite knowledge in students to inform lesson-planning and instructions before beginning the course.
Interpret objective tests	
9 (θ=-0.33; λ=.24)	*Many teachers score classroom tests using a 100-point percent correct scale. In general, what does a student's score of 90 on such a scale mean?* The student answered 90% of the items on this test correctly.
Interpret standardized score	
16 (θ=-0.22; λ=.31)	*John got 7 in all three English courses, Integrated Skills of English, English Speaking, and Writing. The scores are all Stanine scores (which is a method of scaling test scores on a nine-point standard scale with a mean of five and a standard deviation of two). Which of the following is a valid interpretation of this score report?* John had the same percentile rank on the three tests.
Engage students in assessment	
22 (θ=0.21; λ=.41)	*Suppose that you want your students to review each other's argumentative essays. Which of the following can maximize the peer review outcomes?* Use one argumentative essay written by a previous student, and ask students to discuss about the strengths and weaknesses. Develop a rubric with students and ask them to grade accordingly.
Identify unethical assessment practices	
19 (θ=1.32; λ=.24)	*Which teacher's action is considered ethical?* Teacher D asked students to give a group presentation based on the given topic, and gave same grades to every group member.

RQ3: Do teachers' demographic characteristics (i.e., gender, age, professional title, qualification, years of teaching, assessment training, university level, and region) have an effect on their AL performance?

The 3PL IRT θ score for each person was treated as the dependent variable in a multi-factorial analysis of variance. The model tested simultaneously main effects for gender, years of teaching, qualification, title, region, assessment training experience, and level of university along with all two-way interactions. Although the model was statistically significant (F=2.225, p<.001, power=1.00), only those main and interaction effects that were simultaneously statistically significant p<.05 and sufficiently powerful (i.e., >.80) are reported. On that basis, no statistically significant results were found for any of the main effects. This means that no single demographic characteristic had a significant impact on teachers' AL performance. In contrast, statistically significant results were found only for two interaction effects (i.e., region by university $F_{(8, 882)} = 2.123$, $p = 0.03$, $d = 0.29$, power = 0.85; qualification by university $F_{(4, 886)} = 3.171$, $p = 0.01$, $d = 0.26$, power = 0.82). However, since the practical significance of these results shown in Cohen's d was small, it seems best to conclude that teachers' demographic characteristics had little influence upon their AL.

Discussion

TALQ validation: LAL constructs and inventories needed

The purpose of our psychometric properties analyses of the adapted TALQ items was addressing the pressing need for investigating the validity and reliability of this AL measure (Gotch & French, 2014). In contrast to earlier studies (Plake, et al. 1993; Zheng, 2010), which used the classical test theory approach of summing the number of items scored correctly, the present study provides a more robust estimation of the participants' AL level by removing items with high guessing probability, low factor loading and negative discrimination index. The psychometric analyses confirm that there is validity only for a limited set of items, consistent with Fulcher's (2012) skepticism over the constructs in the TALQ for having 'little operational structural integrity' (p. 117).

We infer that the items and the underlying constructs in the TALQ may not be a good representation of AL for Chinese university English teachers, insofar as the content of the adapted version is aligned with the original. Given that teacher AL is subject to the changes in contexts, policy, and culture (Xu & Brown, 2016), our finding suggests that adapting the imported TALQ which was designed in the U.S. context 30 years ago at the surface level is insufficient for measuring teacher AL in the setting of contemporary

Chinese higher education. In other words, AL items need to be revised at a deeper level or even rewritten to better align with the Chinese assessment contexts. The evidence that the newly added items (i.e. items 22, 23, 24) work well with the sample further confirms that items that are contextually grounded would be more likely to be valid, highlighting the importance of alignment among constructs, items and contexts. It is thus suggested that AL measures need to incorporate both generic assessment knowledge (e.g., reliability, validity) applicable to all contexts and contextually-grounded principles, values and traditions. It also needs to differentiate educational settings because the assessment priorities differ from primary through secondary to tertiary levels.

Second, teacher AL might not be multi-dimensional as the TALQ originally intended, as the multi-dimensional model was rejected by our data. Although it is not yet clear concerning how the dimensionalities of teacher AL are operationalized, it is likely that the items from the TALQ needs rewriting given that many items have high guessing values. A close inspection of the items and the underlying constructs reveals that the constructs themselves might be problematic. For example, the acts of 'administering, scoring, and interpreting assessment results' were aggregated into one single standard, while the required competencies for each of these acts are distinct from one another. It suggests that using the *Standards* as a blueprint for developing AL measure may have underscored some of the important competencies, and that developing survey tests of teacher AL needs to build upon more recent professional standards (e.g., JCSEE, 2015) and contemporary assessment policies in specific contexts.

Having said that, the most challenging task for language assessment researchers is to develop "a grounded language assessment inventory" (Inbar-Lourie, 2013, p.6). The reasons for the unavailability of such a measure are unclear, but we believe that the prerequisite for creating one is to understand the nature and extent of the language components within the LAL constructs (Inbar-Lourie, 2013), in particular how these constructs might overlap with those in AL (Taylor, 2013). Three approaches to achieving these goals are suggested. First is a top-down approach driven by the need to fully understand prescribed views of what should be included in LAL. Comparative studies of the content of language assessment textbooks with that in educational assessment textbooks would be helpful to identify which components are generic concepts and which are 'bolt-on' components from language assessment. Additionally, it would be equally helpful to develop standards for language teacher assessment practice by accommodating codes of practice that were originally developed for language testing professionals, such as *Codes of Ethics* (ILTA, 2001) and *Guidelines for Practice* (ILTA, 2007), to meet teachers' needs in their assessment practice. Second is a bottom-up approach that investigates teachers' perspectives concerning the assessment competency or skills that

they believe are needed for good practice in language pedagogy is warranted. As assessment literacy may vary in accordance with curriculum goals, school settings, and socio-political contexts, either local or international studies will be helpful for generating essential components for LAL driven by practical needs (e.g., Vogt & Tsagari, 2014). Third is a situated approach based upon close observations of language teachers' assessment practices. It will help identify whether and how nuanced differences or progressive stages of LAL exist. Particularly, differentiating between competencies needed respectively for summative and formative assessment will help define LAL in a more effective way (Leung & Rea-Dickins, 2007). These three approaches may jointly establish solid LAL constructs, which will lay the groundwork for developing an LAL measure.

A basic level of AL in certain dimensions: why and what are we left with?

Although our claim of the teachers' basic level of AL on specific dimensions cannot be generalized to an overall AL level, it can be inferred that this basic level in certain AL dimensions is insufficient for university English teachers to take on their enormous assessment responsibilities. This adds to the bulk of international evidence that teachers' AL knowledge is adequate (Alkharusi et al., 2011; Mertler, 2004; Plake et al., 1993), suggesting that assessment illiteracy is a global concern.

Three factors may account for this insufficiency. First is a lack of assessment policies and professional standards acting as quality assurance for teacher assessment practice. As noted earlier, only the CECR (CMoE, 2007) briefly describes the parallel position of formative and summative assessments, and there are no other authoritative documents prescribing standards for teacher assessment practice. Without standards to compare against, teachers themselves cannot judge whether they are making valid and reliable judgments in their assessment practice. Likewise, administrators and teacher educators do not have guidelines to follow to evaluate teacher AL. Second is the absence of AL standards in recruitment criteria for university English teachers in China. Given that no threshold level of AL is required, teacher AL cannot be guaranteed in the first place. Only when AL is considered as part of essential competency in teacher recruitment can a basic AL level be guaranteed. Third is inadequate assessment training in pre- and in-service teacher education programs. Since the adapted TALQ is basically a test of assessment principles, the barely satisfactory result is justifiable due to an absence of assessment input from formal and informal learning experiences as reported by the respondents.

These policy and curriculum conundrums, regrettably, are beyond teachers' immediate control. Instead, policymakers, language assessment specialists, teacher educators, and university administrators ought to jointly solve these problems. First, policy makers need to devise assessment policies, professional standards and guidelines, and codes of ethics,

all of which can be used as reference for teacher accreditation and licensure. This is the first step to, and an essential condition of, ensuring teacher AL. Second, university administrators ought to consider teacher AL as one of their recruitment criteria and devise concrete quality assurance mechanism to maintain the professional standards. They should also include AL into constant teacher evaluation systems and use the evaluation results for actions such as promotion, salary bonuses, and employment. Third, pre-service teacher education programs need to give curricular prominence to AL through systematic assessment courses; while in-service programs can take advantage of technology-facilitated resources (e.g., online tutorials or webinars) to make assessment training more accessible to busy teachers.

Impact on teacher AL: assessment training, professional experience and institutional context

The findings of the weak ability of the participants' personal and employment demographic factors to explain their AL corroborate some prior studies (e.g. Brown, 2008; Zhang & Burry-Stock, 2003) while contradicting others (cf. DeLuca, Chavez, & Cao, 2013; Graham, 2005; Lukin et al., 2004), adding to the controversy of the literature. For example, the finding that assessment training experience had no effect on participants' AL is consistent with Brown's (2008) but not others' (Alkharusi, et al. 2011; Mertler, 2009) which reported increased AL with intensive assessment training. While this finding seems discouraging to teacher educators, it could be due to the fact that the forms of assessment training were not specified in the items. Alternatively, the effects of assessment training may have faded due to the relatively long term (>7 years) of teaching service that the majority (>700) of participants had. To keep in-service teachers well informed of assessment principles, sustainable assessment training programs need to be developed and executed. Workplace-based assessment training tailored to teachers' needs and their institutional contexts would be particularly helpful. Future research into the impact of assessment training on teacher AL needs to delineate the influence of each training strategy, as well as the mechanism to determine the appropriateness of these strategies in particular contexts.

Different from findings of earlier studies (cf. Hoover, 2009; King, 2010), the small effect of teachers' professional experience (i.e., years of teaching, professional titles, and qualifications) could be interpreted in the following three ways. First, it may result from an absence of quality assurance for teachers' assessment practice in China. Given that these teachers received little assessment training in their pre- and in-service teacher education and that AL is not one of the threshold criteria for the profession, it would make no difference to AL whether one has accumulated more teaching experience, moved upward along professional trajectory, or secured a higher degree. Second, it may

suggest the relatively independent attribute of AL. That is, AL may be less associated with these professional categorizations of teachers, but more related to other unexplored factors, such as complex compromises that teachers need to make in their assessment practices (Xu & Brown, 2016). To uncover these hidden factors, ethnographic or longitudinal studies would be particularly helpful. Third, the similarity of responding across these factors points to a community of understanding, at least among university English teachers in China. This suggests that, without substantial changes to contextual factors, increased AL is unlikely to occur.

Another finding to note is institutional context as a potentially important factor for teacher AL, as university type had a role in both of the interaction effects (i.e., university * region, university * qualification). Although the effects were small, it suggests that the potential impact of institutions on teacher AL can be multiple. We infer that first-tier universities which usually have a higher standard of staff recruitment are more likely to have quality assurance of teacher assessment practice and assessment training, compared to their counterparts in the lower tiers. Future studies need to be directed to find out what specific institutional factors are exerting the influence. However, as influences from these factors may be subtle yet persistent, it would not be easily detected by quantitative measures. Comparative studies in different workplace sites would be particularly illuminating to understand what and how institutional contexts play a role in shaping teacher AL.

Conclusion

As the first study to subject an adapted version of the TALQ to rigorous psychometric analyses and to administer the measurement among university English teachers in China, our central finding was that these teachers have a very basic level of AL in certain dimensions with limited influence from demographic characteristics. It contributes to AL/LAL research with evidence from university English teachers in China corroborating the conclusion that teacher AL is insufficient and needs development.

Our psychometric analyses of the adapted TALQ point out that cross-context AL measures may be impossible, as assessment principles need to be operationalized in local classrooms. It suggests a need for substantial revisions or rewriting of AL measures to keep abreast of current professional standards and recent educational assessment research. New AL measures need to be contextually-grounded based upon the assessment policy, values and traditions within the specific context (e.g., DeLuca, LaPointe-McEwan & Luhanga, 2016). For language teacher assessment literacy, LAL constructs need to be specified before an LAL measure could be developed.

This study has implications for policy and practice of language teacher assessment education in China and elsewhere in the world. It indicates a pressing need for establishing professional standards and guidelines for language teacher assessment literacy and taking it into consideration of teacher licensure on the national level to ensure high quality assessment practices. It also calls for sustainable AL enhancement programs throughout teacher professional life and joint efforts from university administrators, teacher educators, and teachers.

Acknowledgement

Financial support for this study was provided in part by the National Social Science Foundation of China (grant #12CYY026), Doctoral Dissertation Grant from the International Research Foundation for English Language Education (TIRF), Assessment Research Award from British Council, TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment from Educational Testing Service, and Postgraduate Studentship from the Faculty of Education, The University of Hong Kong. The support of Prof. Jun Liu from Stony Brook University in data collection is acknowledged, as well as the special help of Ms. Renxia Zhang in preparing the online survey. The advice from Prof. David Carless from The University of Hong Kong is especially appreciated. Special thanks go to all the teachers who voluntarily participated in the study, as well as the data analysis assistance provided by The Intelligent Algorithm and Intelligent Software Studio with their website (http://www.autosem.net/).

References

Alkharusi, H. (2011). Psychometric properties of the teacher assessment literacy questionnaire for preservice teachers in Oman. *Procedia - Social and Behavioral Sciences, 29*(0), 1614-1624.

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

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

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

Brown, G. T. L. (2008). Assessment literacy training and teachers' conceptions of assessment. In C. Rubie-Davies & C. Rawlinson (Eds.), *Challenging Thinking about Teaching and Learning* (pp. 285-302). New York: Nova Science.

Campbell, C., & Evans, J. A. (2000). Investigation of preservice teachers' classroom assessment practices during student teaching. *The Journal of Educational Research, 93*, 350-355.

Chen, Q., May, L., Klenowski, V., & Kettle, M. (2014). The enactment of formative assessment in English language classrooms in two Chinese universities: Teacher and student responses. *Assessment in Education: Principles, Policy & Practice, 21*(3), 271-285.

CMoE. (2007). *College English Curriculum Requirements*. Beijing: Foreign Language Teaching and Research Press.

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. New York Guilford Publications.

DeLuca, C., & Bellara, A. (2013). The current state of assessment education: Aligning policy, standards, and teacher education curriculum. *Journal of Teacher Education, 64*(4), 356-372.

DeLuca, C., Chavez, T., & Cao, C. (2013). Establishing a foundation for valid teacher judgement on student learning: The role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice, 20*(1), 107-126.

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

DeLuca, C., LaPointe-McEwan, D., & Luhanga, U. (2016). Approaches to classroom assessment literacy: A new instrument to support teacher assessment literacy. *Educational Assessment, 21*(4), 248-266.

Dillman, D. A. (1991). The design and administration of mail surveys. *Annual review of sociology, 17*, 225-249.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavior Psychology, 70*(1), 113-136.

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

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

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice*, 33(2), 14-18.

Hoover, N. R. (2009). *A descriptive study of teachers' instructional use of student assessment data.* Unpublished doctoral dissertation. Virginia Commonwealth University, Richmond, VA.

International Language Testing Association (ILTA). (2001). *Code of ethics*. Retrived March 3rd, 2016, from www.iltaonline.com/images/pdfs/ILTA_Code.pdf

International Language Testing Association (ILTA). (2007). *Guidelines for practice*. Retrieved March 3rd, 2016, from www.iltaonline.com/index.php/enUS/?option=com_content&view=article&id=122&Itemid=133

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

Inbar-Lourie, O. (2013). Language assessment literacy. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp.1-9). Blackwell Publishing Ltd.

Joint Committee on Standards for Education Evaluation (JCSEE). (2015). Classroom assessment standards for PreK-12 teachers. [Kindle version]. Retrieved from https://www.amazon.com/Classroom-Assessment-Standards-PreK-12-Teachers-ebook/dp/B00V6C9RVO

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

King, J. D. (2010). *Criterion-referenced assessment literacy of educators*. Unpublished doctoral dissertation. University of Southern Mississippi, Hattiesburg, MS.

Leung, C., & Rea-Dickins, P. (2007). Teacher assessment as policy instrument: Contradictions and capacities. *Language Assessment Quarterly, 4*(1), 6-36.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice, 23*(2), 26-32.

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

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

Mertler, C. A., & Campbell, C. S. (2005). Measuring tachers' knowledge and application of classroom assessment concepts: development of the *Assessment Literacy*

Inventory. Paper presented at the anuual meeting of the American Educational Research Association, Montreal, Quebec, Canada, April.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). Old Tappan, NJ: MacMillan.

Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing, 30*(3), 381-402.

Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educaitonal Measurement: Issues and Practice, 12*(4), 10-12,39.

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

Raykov, T., & Marcoulides, G. A. (2007). *A first course in structural equation modeling*. New York: Psychology Press.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1-36.

Rudner, L. M. (2012). PARAM Calibration Software for the 3 Parameter Logistic IRT Model (freeware). (Version 0.93). Retrieved on Dec 15, 2014, from http://echo.edres.org:8080/irt/param/

Schafer, W. D., & Lizzitz, R. W. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education, 38*(3), 57-63.

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

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-245.

Stiggins, R. J. (2010). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 233-250). New York, NY: Taylor & Francis.

Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing, 30*(3), 403-412.

Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly, 11*(4), 374-402.

Wang, S. & Wang, H. (2011) On the state of college English teahcing in China and its future development. *Foreign Languages in China, 5*, 4-11, 17.

Wilson, E. B., & Hilferty, M. M. (1931). The distributions of chi-square. *Proceedings of the National Academy of Sciences, 17*(12), 684-688.

Xie, H. (Ed.). (2014). *Education Statistics Yearbook of China*. Beijing: People's Education Publishing Hourse.

Xu, Y., & Brown, G.T.L. (2016). Teacher assessment literacy in practice: A

reconceptualization. *Teaching and Teacher Education*, 58, 149-162.

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*, 323-342.

Zheng, D. (2010). An investigation into the assessment literacy of secondary and primary school teachers: A report from Z province. *Global Education Review 39* (2), 31-36.