Language Testing in Asia

RESEARCH Open Access

Toward measuring language teachers' assessment knowledge: development and validation of Language Assessment Knowledge Scale (LAKS)

Elçin Ölmezer-Öztürk* and Belgin Aydin

* Correspondence:
elcinolmezerozturk@anadolu.edu.tr
Department of Foreign Language Education, Anadolu University, 26470 Eskişehir, Turkey

Abstract

This study reports on the development and validation of Language Assessment Knowledge Scale, LAKS, which aims to measure teachers' language assessment knowledge. The initial development of the scale included a thorough literature review, creating an item pool, and an initial expert opinion stage. The later process, which led to the development of the final form, consisted of several detailed stages including organizing meetings with practitioners in the field, gathering opinions of experts in language testing and assessment, and piloting the scale. At the end of this long-lasting validation process, LAKS with 60 items and 4 constructs (assessing reading, assessing listening, assessing writing, and assessing speaking) was completed by 542 EFL teachers working in higher education context. The statistical procedures during the analysis process included second-order confirmatory factor analysis, and Cronbach alpha for the reliability. The findings revealed that LAKS had a good model-data fit with the obtained factor loads, and Cronbach alpha coefficients were satisfactory. Concluding that LAKS can serve as a valid and reliable instrument to measure language teachers' assessment knowledge, the study offers several suggestions for further research.

Keywords: Language assessment knowledge scale, Language assessment literacy, Developing and validating a scale, Language testing and assessment, EFL teachers

Introduction

Assessment in any education context is mainly associated with learners, yet the teachers' role in determining the result of the assessment and learners' success or failure is undeniable. The literature (Calderhead 1996; Malone 2013) touches upon the critical role of teachers in assessment process and the importance of their decision-making. Research also indicates that teachers are not well equipped with sufficient knowledge on testing and assessment, and they are not ready for their roles as assessors (Mertler 2005).

The discussion regarding the role of teachers in assessment and their competency in this domain has led to the emergence of a relatively popular concept, "assessment literacy." The term was coined by Stiggins (1991), and rooted in mainstream educational and psychology research mostly. According to Stiggins, assessment literate teachers

know "what they are assessing, why they are doing it, how best to assess the skill, knowledge of interest, how to generate good examples of student performance, what can potentially go wrong with the assessment, and how to prevent that from happening" (p. 240).

There are many studies related to assessment literacy of both pre- and in-service teachers in the literature. Teachers' assessment literacy levels have generally been measured through questionnaires. Teacher Assessment Literacy Questionnaire (TALQ), Assessment Literacy Inventory (ALI), Classroom Assessment Literacy Inventory (CALI), and Assessment Literacy Inventory (ALI) can be listed as some of the popular instruments. TALQ, developed by Impara, Plake, and Fager in 1993 revealed a mean score of 23 out of 35 of in-service elementary and secondary school teachers in the USA. ALI, the renamed version of TALQ, by Campbell, Murphy, and Holt in 2002 focused on pre-service teachers' assessment literacy level and found out that even after completing a course on educational assessment, pre-service teachers got a mean score of 21. CALI, developed by Mertler in 2005, compared the assessment literacy levels of in-service and pre-service teachers. The results revealed higher literacy levels for in-service teachers. The last instrument developed by Mertler and Campbell (2005) was Assessment Literacy Inventory (ALI), and the items were developed considering the "Standards for Teacher Competence in Educational Assessment of Students" (AFT, NCME, and NEA, 1990). The results indicated that the mean score of the respondents was 23.83 out of 35, meaning that the pre-service teachers in this study had a low level of assessment literacy. As seen in all these studies, although in-service teachers had a higher level compared to pre-service teachers, both groups had low literacy levels of assessment literacy.

Taking the definition of assessment literacy as the root and focusing on the competency of language teachers, a novel term has flourished which is language assessment literacy (LAL), and there exist several definitions of language assessment literacy in the field. According to Taylor (2009, p. 24), language assessment literacy, in general sense, is "the level of knowledge, skills and understanding of assessment principles and practice that is increasingly required by other test stakeholder groups, depending on their needs and contexts." Malone also (Malone 2013, p. 329) defined it as "language teachers' familiarity with testing definitions and the application of this knowledge to classroom practices in general and specifically to issues related to assessing language". Inbar-Lourie (2017) stated that language assessment literacy requires additional competencies when compared to assessment literacy and added that it is the combination of assessment literacy skills and language-specific skills.

Although it is stated to be a relatively new field by Fulcher (2012), there has been an increasing research trend on language assessment literacy of teachers. For instance, Lam (2015) carried out a study to investigate the overall language assessment training in five Hong Kong institutions, and more specifically aimed to find out how two language assessment courses facilitated or inhibited the language assessment literacy of pre-service teachers. The analysis of the programmes showed that there was insufficient support to foster LAL, and the training for LAL was inadequate. Based on the perceptions of the participants, three themes came out which were perceptions of LAL in an examination-oriented culture, experience of course-based language assessment training, and restricted application of LAL in authentic school contexts. In another study, Tsagari and Vogt (2017) carried out a mixed-design study covering both quantitative and

qualitative data. The results demonstrated that the participants teachers' perceived LAL was not sufficient, and they did not feel themselves prepared effectively for assessment-related practices. Additionally, it was found that teacher education programmes were not giving the efficient and sufficient education and training in language assessment to the pre-service teachers. The aim was to find out the teachers' perceptions of LAL and their individual needs related to language testing and assessment. In another recent study by Baker and Riches (2017), the LAL development of the teachers was examined. One hundred twenty Haitian high school teachers participated in the study, and the data were collected via feedback on drafts of revised exams, survey with teachers, and teacher interviews. Some workshops were designed in 2013 for the participants, and this study took these workshops as its basis. It was concluded that LAL development of the teachers was clear after these workshops, and the main areas where the teachers' LAL levels increased were creating reading comprehension questions, integrating vocabulary task, basing all exam sections on the same topic, increased attention of the connection between teaching and assessment, broadening of the teachers' understanding of the construct of language ability, teachers' beliefs concerning their supportive role, and finally learning about reliability, validity, and practicality.

Purpose of the study

When the studies in the literature related to language assessment literacy are investigated, it is seen that they mostly focus on the needs and perceptions of language teachers in terms of testing and assessment (Fulcher 2012; Inbar-Lourie 2008; Malone 2013; Scarino 2013; Lam 2015; Baker and Riches 2017) and the main conclusion is that teachers should be supported and encouraged to be more assessment literate. However, the first step toward encouraging language teachers to be more assessment literate is to determine what teachers know and do not know regarding language assessment. Though there has been an increasing research tendency toward language assessment literacy of teachers, what is primarily needed as a starting point is the investigation of language assessment knowledge of EFL teachers, because assessment knowledge is the core of assessment literacy.

Additionally, the research instruments in the literature are not specifically designed to measure language assessment knowledge of EFL teachers. The measurement tools mentioned in the previous section (Impara et al. 1993; Campbell et al. 2002; Mertler and Campbell 2005) were developed in the light of the "Standards for Teacher Competence in Educational Assessment of Students" (AFT, NCME, and NEA 1990). The items in these measurement tools were based on seven standards; thus, seven different features were assessed through these tools and assessment literacy of the participants was assessed via multiple choice items in these tools. Despite this advantage, the items in these measurement tools are general and applicable to all fields related to educational assessment. In other words, the items do not change according to the teachers whose majors are Maths or English. Hence, it can be concluded that the items are not field-specific. Even if these items are answered by EFL in-service or pre-service teachers, what is concluded is the general assessment knowledge of these teachers. The scores they get from these tools indicate their assessment knowledge related to general education, not language-related knowledge.

Apart from these measurement tools, there are also some tools aiming to measure language assessment literacy of EFL teachers. One belongs to Tao (2014) who developed four distinct scales one of which aims to measure classroom assessment knowledge of EFL teachers. This tool includes some elements related to language, but it still has too many items related to general education. Additionally, the items are not divided into language skills. The next one belongs to Kremmel and Harding (Towards a comprehensive, empirical model of language assessment literacy, forthcoming) whose instrument is a five-point Likert scale. In their scale, participants have to choose one of the options ranging from not knowledgeable/not skilled to extremely knowledgeable/extremely skilled. Thus, these items were developed to find out the needs of EFL teachers; but, as these are self-reports, what is concluded relies on the perceptions or opinions of the participants. The last one is Vogt and Tsagari (2014)'s study in which they aimed to find out whether the participants were trained or needed training in various domains related to assessment such as classroom-focused LTA or purposes of testing. Similar to Kremmel and Harding, this tool also seeks for self-reports of the participants, and aims to measure their language assessment knowledge and practices based on a perception level.

For these two major reasons, there is an urgent need for the investigation of this knowledge, and this study aims to develop and validate Language Assessment Knowledge Scale (LAKS) as a new and to-the-point instrument to be used for that purpose. In this sense, this paper aims to address to the following research question throughout the study.

1. What are the psychometric properties of language assessment knowledge scale?

Methodology

The research context

In Turkey, preparatory programs at universities provide intensive foreign language education to university students before they start their education in their academic fields. Students who finish these programs successfully by getting the required grades from the proficiency exams can continue their education in their departments. Otherwise, they have a chance to get the language education for one more year until they get proficient. In some cases, students who do not have English-medium instruction in their departments can prefer to learn a foreign language voluntarily by attending these intensive language education programs.

In English preparatory programs, all educational and instructional practices are conducted by language teachers. They are the graduates of English language-related departments such as ELT, English language and literature, English linguistics, etc. and are recruited based on their scores from nation-wide exams and interviews carried out by the school administrators. In addition to their weekly course load about 15–20 h, they are also responsible for some other office works such as curriculum and material development, professional training and testing, and assessment practices. These programs of the universities are one of the workplaces where teachers are expected to teach English, and assess their learners in each skill. This context was purposefully selected for this study because they are the contexts in which each language skill is given importance, and as a result, all skills are assessed. The assessors in these preparatory programs are language teachers. The problem is that language teachers are responsible for all the assessment-related activities in preparatory programs, but how knowledgeable or

competent they are in assessing their learners is open to discussion (Hatipoğlu 2015). As a starting point, language assessment knowledge of language teachers should be determined. This identification is vital because by detecting the strengths and weaknesses of language teachers, the needs of language teachers could be specified. Based on these needs, testing and assessment course in pre-service education and teacher professional development programs related to language assessment can be designed and developed.

Participants

The participants of the study included 542 teachers working at English preparatory programs at various universities in Turkey. The scale was prepared in an online format and sent to almost all universities having a preparatory program. After a 1-month period, the number of the teachers who fully completed the scale was 542 and the demographic features with the numbers are presented in the following Table 1.

Developing the items

First of all, in order to provide a deep theoretical background to the instrument, the books referenced so far on language testing and assessment (Harris 1969; Hughes 1989; Heaton 1990; Bachman 1990; Bailey 1998; Alderson 2000; Buck 2001; Weigle 2002; Brown 2003; Luoma 2004; Fulcher and Davidson 2007, 2012; Coombe et al. 2012, etc.) were read by the researcher. While reading, all the knowledge elements which were stated in those books as "need-to-know" about testing or assessing language skills; that is, reading, listening, writing, and speaking, were listed by the researcher. Then, the researcher chose the ones repeated in references mentioned above for the item pool for each language skill. This list consisted of 237 items in total (49 items for reading, 61

Table 1 Participants and the demographic features

Demographic feature	Number of participants	Percentage
Gender	Male—174	32
	Female—368	68
Years of experience	1–5 years—86	16
	6–10 years—173	32
	11–15 years—114	21
	16–20 years—100	18
	More than 21—69	13
Educational background	BA—238	44
	MA—255	47
	PhD—49	9
The BA program graduated	ELT—347	64
	Non-ELT—195	36
The current workplace	State University—372	68
	Private University—170	32
Had a separate testing/assessment course in pre-service	Yes—282	52
	No—260	48
Ever been a member of a testing office	Yes—260	Yes – % 48
	No—282	No – % 52

for listening, 74 for writing, and 53 for speaking). Next, three experts with a PhD degree went over the item pool in detail focusing carefully on the comprehensibility and orthography of the items and the compatibility of each item for the language skill it was listed in. At the end of this initial step, 17 items (3 items from reading, 2 from listening, 6 from writing, and 6 from speaking) were removed from the instrument and the very initial format of the scale had four constructs; assessing reading (46 items), assessing listening (59 items), assessing writing (68 items), and assessing speaking (47 items), consisting 220 items in total.

The process of ensuring content validity

At the second stage, individual meetings with ten teachers having various years of teaching experience and educational background from preparatory programs of different universities were held. In these individual meetings, the teachers were asked to read the items and make comments on whether the items were clear to them and they had any difficulty in understanding the terminology in the items. At the end of those meetings, no item was removed from the list but several revisions were made based on the suggestions provided by the teachers to make the wordings clearer for further stages.

The third stage of developing the instrument included getting the opinions of experts in the field of ELT. For this, the instrument was designed in a format having four different parts, each for a different language skill and the items were listed in these parts. For each item, the researcher put three choices as "necessary, not necessary, needs revision (please justify)" similar to a Likert scale and the items were provided to the experts, 14 academicians who studied on testing and assessment or gave related courses in higher education level in the fields of English language teaching and testing and evaluation at different universities. In 1 month, 11 of the experts responded to the initial format of the instrument and provided feedback on each item. Based on the suggestions provided by these experts, 67 items were removed from the instrument, and revisions were made on several items. At the end of all these stages, 153 items remained in the scale (reading 37 items, listening 33 items, writing 48 items, and speaking 35 items).

At the fourth stage of the process, the scale was presented to real practitioners, which was believed to contribute significantly to the validation of the instrument; the researcher organized a meeting with the English preparatory program-testing office of one of the leading universities in Turkey. The meeting included 18 teachers; 6 of them had PhDs or MAs in testing and evaluation or in ELT. They were sent the instrument before the meeting and were asked to respond and comment on it beforehand. The meeting in which the participants and the researcher discussed the validity, comprehensibility, and compatibility of each and every item lasted about 5 h. At the end of the meeting which provided the researcher a deeper insight from the perspectives of the practitioners, 41 items were removed from the instrument, and several revisions were made on the remaining ones. Finally, the instrument which is called Language Assessment Knowledge Scale (LAKS, henceforth) consisting of 112 items (reading 28 items, listening 26 items, writing 34 items, and speaking 24 items) were ready for the piloting process. This removal and/or revision process in all these stages is shown in Table 2.

The final version including 112 item were piloted with 50 teachers who were then excluded from the actual study. They were asked to both complete the scale and make

Table 2 The number of removed items throughout the validation process

	Reading	Listening	Writing	Speaking	In total
1st stage (three experts with PhD in ELT checking for comprehensibility)	49	61	74	53	237
	− 3	− 2	− 6	− 6	− 17
2nd stage (Checking with 10 teachers)	46	59	68	47	220
	–	–	–	–	–
3rd stage (Checking with academicians)	46	59	68	47	220
	− 9	− 26	− 20	− 12	− 67
4th stage (Training with testing office members)	37	33	48	35	153
	− 9	− 7	− 14	− 11	− 41
5th stage (Piloting with 50 teachers and expert opinion)	28	26	34	24	112
	− 13	− 11	− 19	− 9	− 52
The final version	15	15	15	15	60

comments on it. However, after receiving their answers, it was observed that the participants tended to give the same answers (all true or all false) toward the end of the scale and some of the participants did not even finish completing. Furthermore, the comments made by those participants revealed that there were too many items to respond in the scale and it took too much time, demotivating them to complete. Based on this feedback, which consisted of elements that had the potential to influence the validity and the reliability of the scale negatively, five academicians who were experts in foreign language teaching and assessment had long discussions on each item in the instrument and decided to keep the items that were fundamental for a language teacher to know regarding the assessment of a foreign language, and the other items were eliminated from the study. At the end, the remaining 60 items, with 15 in each construct were sent to all the language teachers working at the preparatory programs in Turkey in an online platform.

Data collection procedure

The data of this study were collected during the early days of the spring semester of 2017–2018 academic year. Among 122 universities in Turkey (85 state and 37 private), the scale was sent to the ones with English preparatory programs. Among these universities, which were decided as the context of the study, 37 state and 18 private contributed to the data collection process of the study. The scale was sent in an online format to all the teachers working in these programs. During data collection process, reminder e-mails were sent to the participants and the head of their schools by the researcher in order to encourage the participants to respond to the scale. This process lasted about one and a half month and at the end of this period, 542 participants responded LAKS completely, and these participants formed the core data of this study.

Findings

The research question of this study aimed to reveal the psychometric properties of LAKS. First, in order to confirm the compatibility of the items with the constructs (assessing reading, assessing listening, assessing writing, and assessing speaking), and

the compatibility of these constructs with language assessment knowledge; in other words, the model-data fit in general, second-order CFA was performed using the Mplus 7.0 package program. Since the responses given for each item were categorical, WLSMV was used as the proficiency estimator. Since CFA is included in the structural equation modeling family, the model data compatibility was first investigated for the results of CFA. The results and interpretations are as follows.

In the structural equation modeling studies, the expected chi-square value is not significant, in other words, the value of "*p*" must be bigger than .05. However, this value can be misleading because it is sensitive to the size of the sample. For this reason, the value obtained by dividing the chi-square by degrees of freedom is generally reported. At this point, the value of the model which is below 2.5 indicates a good fit. Besides, in second-order CFA, the RMSEA values smaller than .08 indicate a good one. Moreover, good fit for CFI and TLI values is for values above .90 and .95 respectively (Li-tze and Bentler 1999; Byrne 2012; Çokluk et al. 2012). At this point, the values presented in Table 3 (.028 as the RMSEA value, .981 as the CFI, and .980 as the TLI value) revealed a good fit in this study based on the cut-off points. Thus, it can be said that the complete statistics obtained are indicative of a good model-data fit.

Standardized values in the structural equation modeling are interpreted as standardized coefficients in the regression. In the context of CFA, these values are seen as factor loads. Factor loads for each item, standard errors, and *t* values for these values are presented in the Table 4.

The values given in the first column above are referred as standardized path coefficients, and these values are accepted as factor loadings in CFA. The coefficients are valued between − 1 and + 1, and the higher the value is, the higher its relationship with the latent variable is. The second column refers to the standard error values and the third column includes the *t* values, which are obtained by dividing the factor loading of an item to its standard error. Getting higher *t* values increases the significance of the items. The last column gives the R-square values which is equal to the square of factor loadings. This value is between 0 and 1, and as it gets closer to 1, the amount of variance explained in the observed variable increases. Based on these explanations, it can be seen that the factor loadings of most of the items in assessing reading and assessing listening are significant and satisfactory, whereas there exists several items with low factor loadings in assessing writing and assessing speaking.

In the next step, the structural values obtained were reported on the model. This figure is shown below (Fig. 1).

Based on the second-order CFA, the figure above reveals how LAK explains its constructs (assessing reading, assessing listening, assessing writing, assessing speaking) in terms of their variance. Firstly, standardized path coefficients were reported as .98 for assessing listening, .99 for assessing reading, .89 for assessing writing, and .98 for

Table 3 Model-fit indices derived from second order CFA

Fit indice	Cut off for good fit	Value	Comment
Chi-square/df	< 2.5	1.41	Good fit
RMSEA	< .08	.028	Good fit
CFI	> .90	.981	Good fit
TLI	> .95	.980	Good fit

Table 4 Factor loadings for each item

Factor	Item no	Factor loading	SE	*t*	R-square
Assessing reading	1	0.869*	0.021	41.055	0.755
	2	0.536*	0.040	13.342	0.287
	3	0.872*	0.019	46.981	0.760
	4	0.777*	0.022	35.885	0.603
	5	0.064	0.058	1.110	0.004
	6	0.753*	0.031	24.012	0.567
	7	0.842*	0.020	41.343	0.708
	8	0.895*	0.022	40.634	0.801
	9	0.632*	0.039	16.402	0.400
	10	0.279*	0.053	5.238	0.078
	11	0.869*	0.019	46.619	0.756
	12	0.805*	0.020	39.951	0.648
	13	0.990*	0.018	55.177	0.980
	14	0.485*	0.047	10.282	0.235
	15	0.855*	0.027	31.767	0.730
Assessing listening	16	0.470*	0.045	10.354	0.221
	17	0.841*	0.021	40.876	0.707
	18	0.824*	0.021	39.164	0.679
	19	0.266*	0.058	4.587	0.071
	20	0.730*	0.024	30.613	0.533
	21	0.680*	0.039	17.548	0.463
	22	0.631*	0.027	23.514	0.398
	23	0.697*	0.024	29.423	0.486
	24	− 0.043	0.059	− 0.728	0.002
	25	0.021	0.059	0.351	0.000
	26	0.902*	0.018	48.991	0.814
	27	0.576*	0.033	17.316	0.332
	28	0.262*	0.052	5.025	0.069
	29	0.660*	0.026	25.784	0.435
	30	0.969*	0.020	48.969	0.940
Assessing writing	31	0.121	0.066	1.847	0.015
	32	0.889*	0.031	28.386	0.791
	33	0.078	0.062	1.267	0.006
	34	− 0.046	0.071	− 0.644	0.002
	35	0.431*	0.054	7.975	0.186
	36	0.013	0.062	0.207	0.000
	37	0.025	0.076	0.330	0.001
	38	− 0.018	0.066	− 0.269	0.000
	39	0.631*	0.045	14.132	0.398
	40	0.585*	0.044	13.245	0.343
	41	− 0.033	0.062	− 0.530	0.001
	42	0.442*	0.051	8.680	0.195
	43	0.085	0.062	1.371	0.007
	44	0.613*	0.046	13.263	0.376

Table 4 Factor loadings for each item *(Continued)*

Factor	Item no	Factor loading	SE	*t*	R-square
	45	0.609*	0.042	14.467	0.371
Assessing speaking	46	0.077	0.059	1.301	0.006
	47	− 0.019	0.067	− 0.290	0.000
	48	0.076	0.061	1.257	0.006
	49	0.479*	0.044	10.964	0.229
	50	− 0.027	0.059	− 0.463	0.001
	51	0.255*	0.061	4.194	0.065
	52	0.684*	0.033	20.660	0.468
	53	0.304*	0.052	5.803	0.092
	54	0.110	0.063	1.759	0.012
	55	0.916*	0.018	50.230	0.839
	56	0.350*	0.053	6.616	0.123
	57	1.020*	0.014	73.086	1.00
	58	0.845*	0.026	32.142	0.715
	59	0.747*	0.033	22.386	0.557
	60	0.039	0.063	0.613	0.002

*$p < .05$

assessing speaking. That means, one standard deviation change in LAK (1.000) would lead to 0.994 standard deviation change in assessing reading, .986 standard deviation change in assessing listening, .891 standard deviation change in assessing writing, and .981 standard deviation change in assessing speaking, all of which are good indicators of variance explanation. In addition to this, the error variance values were found as .027 for assessing listening, .013 for assessing reading, .206 for assessing writing, and .038 for assessing speaking. In other words, these values mean that LAK explains 97% variance of assessing listening, 98% variance of assessing reading, 80% variance of assessing writing, and almost 96% variance of assessing speaking. In short, as all these values suggest, the model presents a perfect model-data fit in terms of explaining LAK and its constructs.

Reliability analysis

In developing and validating measurement instruments, presenting the statistical values related with the reliability is another important factor. The following table gives the Cronbach alpha coefficients of LAKS in total and its sub-constructs (Table 5).

The findings above reveal that Cronbach alpha coefficient of LAKS in total was .91 which is a highly satisfactory value, and it shows that LAKS has a statistically high reliability to be used as a measurement tool. When the table is examined, it is also seen that the Cronbach alpha value for assessing reading appeared to be .88, again referring to a high level of reliability. The Cronbach alpha coefficient for assessing listening sub-construct was obtained as .78, which means that the scale has internal consistency at an acceptable level. The Cronbach alpha coefficient calculated for assessing writing sub-construct was found out as .49. Since the value is below the acceptable limit of .60 (Hulin et al. 2001), this construct resulted in a lower confidence in internal consistency.

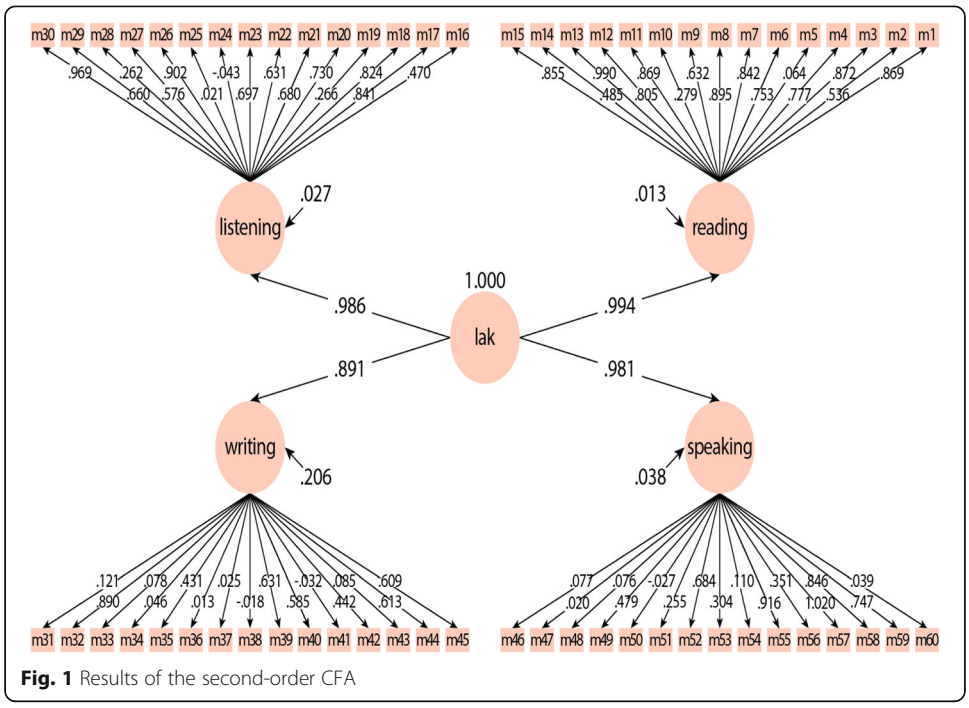

Fig. 1 Results of the second-order CFA

Finally, the Cronbach alpha coefficient for assessing speaking was found to be .65, which is again an acceptable value for the internal consistency.

In addition to the reliability values above, item-total correlations related with each item were also calculated under each skill. The following table presents the values for the items (Table 6).

When the item-total correlation values in the above table are examined, it is seen that items 5 and 10 in reading are relatively low in size. The coefficients obtained for other items in the construct of assessing reading were satisfactory. As for the construct of assessing listening, it is seen that the correlation value obtained for items 9, 10, and 13 in this subtest was relatively low and the other items had satisfactory values. The third construct was assessing writing and in this construct, it is seen that most of the items had low-level correlation values. Finally, the last construct of the scale was assessing speaking, and it is seen that the item-total correlation values obtained for most of the items are below .50 and relatively low in size.

When all the item-total correlation values are examined, it can be concluded that several items under each construct have a relatively low level of item-total correlation. However, after this statistical analysis, three academicians who significantly contributed

Table 5 Reliability analysis for Language Assessment Knowledge Scale (LAKS) and its sub-constructs

Constructs	Cronbach alpha
Language Assessment Knowledge Scale	.91
Assessing reading	.88
Assessing listening	.78
Assessing writing	.49
Assessing speaking	.65

Table 6 Item-total correlation coefficients of the items under each skill

Item no	Item-total correlation			
	Assessing reading	Assessing listening	Assessing writing	Assessing speaking
1	.686	.363	.082	.109
2	.382	.586	.410	− .009
3	.660	.536	.100	.124
4	.564	.204	.009	.282
5	.048	.476	.264	− .037
6	.583	.490	.106	.201
7	.632	.372	.014	.444
8	.720	.449	.106	.176
9	.481	.024	.314	.046
10	.191	.039	.314	.580
11	.699	.625	.040	.284
12	.608	.327	.311	.637
13	.797	.141	.018	.583
14	.312	.439	.235	.484
15	.674	.706	.224	.038

to initial validation process of LAKS were asked to provide expert opinion on those items. Based on their expert opinion, it was decided that these items were important for the content validity of the scale and their contribution to LAKS in general was significant in terms of measuring language teachers' assessment knowledge. Besides, considering model-data fit and reliability coefficients of the constructs, it can be said that the scale presented satisfactory statistical values with those items. Due to all these reasons, the items with relatively low level of item-correlation values were decided to be kept in the scale.

Discussion

After a tough validation process including the opinions of teachers, academicians, and testing practitioners, LAKS was administered to 542 EFL teachers working in higher education contexts. The findings derived from the statistical analysis showed that LAKS had a good model-data fit, and the factor loadings and reliability values were at a satisfactory level. However, writing and speaking constructs had relatively low factor loadings, but they were decided to be kept in the scale due to their high contribution to the content validity.

In its final form (see Additional file 1), LAKS was developed as a scale consisting of four constructs; assessing reading, assessing listening, assessing speaking, and assessing writing, with 15 items for each and 60 items in total. The scale includes the items which specifically refer to the language assessment knowledge of EFL teachers in assessing each skill. In that sense, LAKS makes a unique contribution of the research and practical understanding of language assessment knowledge in the literature. The instruments in the literature focusing on assessment literacy of language teachers either include items all of which are related with general assessment literacy (Impara et al. 1993; Campbell et al. 2002; Mertler and Campbell 2005) or their items require self-reports answers at perception level (Tao 2014; Vogt and Tsagari 2014; Kremmel and Harding: Towards a comprehensive, empirical model of language assessment literacy, forthcoming). By including

items specifically focusing on the knowledge of language teachers on assessing each skill and requiring direct responses from them make LAKS a unique research instrument to measure language assessment knowledge of teachers.

Conclusion

Considering the paucity of research focusing on language assessment literacy and the urgent need for an instrument to measure language assessment knowledge, as the core of language assessment literacy, among language teachers, this study aimed to develop language assessment knowledge scale—LAKS. After a thorough validation process which included literature review, meetings with testing and assessment practitioners, expert opinion, and a piloting process, LAKS with 60 items and 4 constructs (assessing reading, assessing listening, assessing writing, assessing speaking) was completed by 542 EFL teachers working at higher education context. The findings derived from second-order confirmatory factor analysis revealed a good model-data fit. Though some of the items in assessing writing and assessing speaking constructs had low factor loading and item-total correlations, they were not removed from the scale based on the expert opinion considering their significant contribution to the content validity of the scale. Besides, the scale demonstrated satisfactory levels for reliability according to the Cronbach alpha analysis. Based on all the validation and statistical procedures, it can be concluded that LAKS can be used as a valid and reliable instrument to measure EFL teachers' language assessment knowledge.

The major underlying reason behind developing a measurement tool on language assessment knowledge is the urgent need in this field. In other words, in addition to its statistical and procedural validation, LAKS not only presents a baseline for researchers interested in this field but also creates an initial framework to understand assessment knowledge level of language teachers. Besides, this scale might be used not only as the first step to identify the needs of in-service teachers and to develop training programs, it might also have implications for pre-service teacher training programs in preparing future teachers to be more equipped to the field.

A major limitation of this study could be that the scale was developed and validated in Turkish higher education context. Thus, the findings could only be interpreted within the unique features of this context. However, this limitation can also be a suggestion for further research studies. Validation of LAKS in different countries and educational contexts will contribute to our understanding of language assessment literacy of teachers from various groups. Besides, new items that are regarded as crucial in assessing language skills can also be added to LAKS in order to enrich its components and scope. For these reasons, it can be said that validation of LAKS with this study will serve as a baseline for further studies that will focus on language assessment knowledge of teachers in different contexts throughout the world.

Additional file

Additional file 1: Language Assessment Knowledge Scale – LAKS. (DOCX 21 kb)

Acknowledgements

This study is a part of the PhD dissertation entitled as "Developing and validating language assessment knowledge scale (LAKS) and exploring the assessment knowledge of EFL teachers" and completed by Elçin ÖLMEZER-ÖZTÜRK in July 2018.

Funding
No funding was received from any specific funding agencies in the public, commercial, or not-for-profit sectors.

Availability of data and materials
Data and material are available.

Authors' contributions
To conduct the study, EOO and BM worked collaboratively throughout the research process. Both authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 29 June 2018 Accepted: 29 November 2018
Published online: 18 December 2018

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
Bailey, K. M. (1998). *Learning about language assessment: dilemmas, decisions, and directions*. Boston: Heinle & Heinle.
Baker, B. A., & Riches, C. (2017a). The development of EFL examinations in Haiti: collaboration and language assessment literacy development. *Language Testing*. https://doi.org/10.1177/0265532217716732.
Brown, H. D. (2003). *Language assessment: principles and classroom practices*. White Plains, NY: Pearson Longman.
Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.
Byrne, B. M. (2012). *Structural equation modeling with Mplus. Basic concepts, applications, and programming*. New York: Routledge.
Calderhead, J. (1996). Teachers: beliefs and knowledge. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 708–725). New York: Simon & Schuster Macmillian.
Campbell, Y., Murphy, J. A., & Holt, J. K. (2002). Psychometric analysis of an assessment literacy instrument: applicability to preservice teachers. In *Paper presented at the meeting of the Mid-Western Educational Research Association, Colombus, OH.*
Çokluk, Ö., Şekercioğlu, G., & Büyüköztürk, Ş. (2012). *Sosyal bilimler için çok değişkenli istatistik: SPSS ve LISREL uygulamaları*. Ankara: Pegem Akademi.
Coombe, C., Davidson, P., O'Sullivan, B., & Stoynoff, S. (Eds.). (2012). *The Cambridge guide to second language assessment*. Cambridge: Cambridge University Press.
Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly, 9*(2), 113–132.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. London: Routledge.
Fulcher, G., & Davidson, F. (2012). *Routledge handbook of language testing*. London and New York: Routledge.
Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill.
Hatipoğlu, Ç. (2015). English language testing and evaluation (ELTE) training in Turkey: expectations and needs of pre-service English language teachers. *ELT Research Journal, 4*(2), 111–128.
Heaton, J. B. (1990). *Writing English language tests* (2nd ed.). Cambridge: Cambridge University Press.
Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
Hulin, C., Netemeyer, R., & Cudeck, R. (2001). Can a reliability coefficient be too high? *Journal of Consumer Psychology, 10*(1), 55–58.
Impara, J. C., Plake, B. S., & Fager, J. J. (1993). Teachers' assessment background and attitudes toward testing. *Theory Into Practice, 32*(2), 113–117.
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. (2017). Language assessment literacy. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 1–14). New York, NY: Springer.
Lam, R. (2015). Language assessment training in Hong Kong: implications for language assessment literacy. *Language Testing, 32*(2), 169–197.
Li-tze, H., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.
Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.
Malone, M. E. (2013). The essentials of assessment literacy: contrasts between testers and users. *Language Testing, 30*(3), 329–344.
Mertler, A. C. (2005). Secondary teachers' assessment literacy: does classroom experience make a difference? *American Secondary Education, 33*(1), 49–64.
Mertler, C. A., & Campbell, C. (2005). Measuring teachers' knowledge and application of classroom assessment concepts: development of the assessment literacy inventory. In *Paper presented at the annual meeting of the American Research Association, Montreal, Quebec, Canada* Retrieved from https://eric.ed.gov/?id=ED490355.
Scarino, A. (2013). Language assessment literacy as self-awareness: understanding the role of interpretation in assessment and in teacher learning. *Language Testing, 30*(3), 309–327.
Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*, 534–539.

Tao, N. (2014). *Development and validation of classroom assessment literacy scales: English as a foreign language (EFL) teachers in a Cambodian higher education setting*. PhD dissertation. Australia: Victoria University. Retrieved from http://vuir.vu.edu.au/25850/1/Nary%20Tao.pdf

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics, 29*, 21–36.

Tsagari, D., & Vogt, K. (2017). Assessment literacy of foreign language teachers around Europe: research, challenges and future prospects. *Papers in Language Testing and Assessment, 6*(1), 41–64.

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

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.