

Enhancing Language Assessment Skills through Need-based Training for Faculty Members in EFL Tertiary Contexts

Mohd Nazim, Abduljalil Nasr Hazaea

Najran University

Correspondence concerning this article should be addressed to Abduljalil Nasr Hazaea, Department of English Language, Najran University, Saudi Arabia. E-mail: agaleel@gmail.com

In Saudi higher education, assessment has shifted to incorporate intended learning outcomes rather than merely textbook content. Subsequently, faculty members unwillingly participate in high-stakes competitive and harmonized assessment in English as a Foreign Language (EFL) courses during the preparatory year (PY). These challenges emphasize the importance of need-based training for faculty members. Accordingly, this context-specific study scrutinized faculty members' needs as well as the impact of a training program on engaging the participants and on fostering language assessment skills (LASs) among them. In so doing, an action research design used pre- and post-questionnaires and included a training portfolio to collect data from 31 faculty members. The study first identified those needs as instructional skills, design skills, and educational measurement skills. In the context of professional development, the researchers designed a training program based on those reported needs. During training, the participants expressed their satisfaction with the language assessment. After the training, the participants greatly improved their LASs. The paired tests indicated that the faculty members increased their instructional skills, design skills, and skills of educational measurement. Further research is recommended for enhancing LASs among EFL students.

Keywords: language assessment skills, faculty members, preparatory year, high-stakes tests, professional training

Introduction

In Saudi higher education, assessment has shifted to be based on intended learning outcomes rather than only on textbook content. This is specifically for high-stakes tests which determine the career of first year students. In a Saudi university preparatory year (PY), the context of the present study, English as a Foreign Language (EFL) assessment has become harmonized, responsibility is shared, and it is based on learning outcomes. These shifts have led to several reasons for fostering language assessment among faculty members. First, it has become a challenge for many faculty members to design an assessment matrix and a question paper based on learning outcomes. To cope with this challenge, faculty members need new language assessment skills (LASs) after employing content-based assessment for so long. Second, faculty members have different and overlapping roles in harmonized assessment. Their roles can be that of classroom teacher, test writer, course coordinator, and/or exam committee member. These roles make them disengaged and demotivated to participate in language assessment. Third, faculty members joined the English department with various teaching backgrounds, qualifications, and experience levels. Subsequently, the authors experienced poor representations of learning outcome-based assessment in the submitted exam proposals of many test writers. For these reasons, the researchers are interested in conducting a training program for faculty members teaching courses in a Saudi university PY program.

Several studies investigated teachers' language assessment literacy (LAL) in terms of a needs analysis such as Öz and Atay (2017) in Turkey, Lam (2015) in Hong Kong, and Lan and Fan (2019) in China. This is because LAL remains a neglected area in many programs that prepare language teachers (Fulcher, 2012). In fact, LAL involves principles, knowledge, and skills in language assessment. Theoretically speaking, Taylor (2013) reported that classroom language teachers need the highest level of skills. Subsequently, this study puts a focus

on LASs for two reasons. First, faculty members usually need practical solutions when encountering challenges at their workplace. Second, skills coincide with the concept of ‘training’. Under the umbrella of LAL, therefore, LASs are a potential area that no study has concentrated on or addressed in the form of a need-based training program.

At the same time, workplace training is essential for promoting LASs among faculty members to meet 21st century literacies for foreign language assessment in Saudi higher education. Faculty members need to understand certain issues about assessment such as its purpose and outcomes in addition to working with assessment data (Boyd & Donnarumma, 2018). When faculty members are trained in assessment skills, they become more confident with high-stakes tests. It is assumed that the participants in this study will overcome assessment challenges and become more confident aligning learning outcomes and textbook contents after training them on LASs.

Therefore, this context-specific study investigates the effectiveness of need-based training on LASs among faculty members working in a Saudi university PY program. Employing an action research design, this study addresses these three research questions:

1. What is the average level of LAS knowledge among faculty members in a PY EFL context?
2. To what extent are the participants engaged in the training treatment?
3. How effective is professional training on LAS development?

Literature Review

LAL, an emerging area in language assessment, is defined as the knowledge, skills, and principles in language assessment (Fulcher, 2012; Inbar-Lourie, 2008; Pill & Harding, 2013; Taylor, 2013). As a result, several models have been developed to investigate LAL (Brindley, 2001; Fulcher, 2012; Inbar-Lourie, 2008; Pill & Harding, 2013; Taylor, 2013), and how it can enhance the knowledge, skills, and principles in language assessment among various stakeholders (Davies, 2008; Language Assessment Literacy Symposium, 2016). Taylor (2013) summarized a conference on LAL and called for further research in this emerging and challenging field. The conference reported diverse stakeholders including “test writers, classroom teachers, university administrators and professional language testers” (p.410) who need various levels of LAL. As for classroom language teachers, Taylor further showed that they need the highest level of language pedagogy and technical skills. However, they need the lowest level of principles, concepts, and scores and decision making. In other words, language teachers need practical and teaching skills to achieve their assessment duties in a team-work context. On the contrary, the deanship and the program coordinator determine the assessment approach, plan, and matrix. Subsequently, language teachers need the lowest theoretical orientation about language assessment.

Several studies investigated perceptions, principles, and knowledge of assessment among language teachers. Lan and Fan (2019) identified the training needs for assessment among Chinese EFL teachers. Boyd and Donnarumma (2018) concentrated on the principles and perceptions of language teachers in London towards language assessment. Although they conducted a need-based training assessment program, they did not put a focus on LASs. Öz and Atay (2017) examined the perceptions and knowledge of assessment among Turkish EFL instructors in a PY program. The same authors found an imbalance between teachers’ perceptions and classroom reflections. LASs still a great deal of potential for further research. Djoub (2017) advocated teacher training into assessment literacy through conscious professional development courses; something that includes multi-dimensional awareness about the importance, benefits, strategies, and techniques of language assessment. Galichkina (2016) investigated assessment awareness among EFL teachers in Russia through a project-based learning activity. This activity triggered a deep awareness of and confidence about assessment among the participants. The same study recommended integrating this activity in an ESL course. Subsequently, the present study incorporates this activity into its training program.

Gebril and Boraie (2016) designed a training program for EFL teachers in Egypt. They recommended that a similar program be conducted for EFL teachers. Similarly, Lam (2015) explored the training landscape in language assessment among EFL teachers. The study found inadequate language assessment in Hong Kong and

recommended LAL. Hatipoğlu (2015) investigated the issue of language assessment among language teachers in Turkey. The study found a great effect on the assessment culture and assessment experience on pre-service teachers. These findings indicate that changes in assessment practices need well-designed training to cope with new assessment changes. Vogt and Tzagari (2014) identified the training needs in language assessment in seven European countries. They found that EFL teachers need training in language assessment. Instead, teachers compensated for insufficient training by using existing materials for language assessment. Walters (2010) applied ‘standards reverse engineering’ for the sake of critical awareness among ESL teachers and recommended applying it to classroom-based assessment to examine its effect on test construction and use in the classroom. Inbar-Lourie (2008) investigated language assessment courses in language programs to find gaps in teachers’ qualifications. So far, it seems that no study has concentrated on LASs and addressed them through professional development need-based training programs in higher education EFL contexts.

Previous research also called for fostering LASs in the Saudi EFL context (Hakim, 2015; Hazaea, 2019; Hazaea & Tayeb, 2018; Mansory, 2016; Umer, Zakaria, & Alshara, 2018). Hazaea (2019) examined the need for professional development in ELT and recommended further research on LAL. Hakim (2015) suggested the use of training workshops to provide EFL facilitators with a clear and deep consideration of effective LAL. Similarly, Mansory (2016) called for long-term collaborative training and support that would greatly help in raising assessment literacy among teachers. Hazaea and Tayeb (2018) recommended training for EFL teachers to develop LAL. Umer et al. (2018) found a mismatch between Saudi EFL teachers’ assessment tasks and course learning outcomes. Thus, need-based training is critical for promoting LASs among EFL faculty members to meet the 21st century literacies required for foreign language assessment.

Previous research also showed that several questionnaires have been used for LAL. Fulcher (2012) developed a questionnaire to find out the training needs regarding assessment among language teachers. Herrera, Macías, and Fernando (2015) developed a questionnaire for EFL teachers on assessment literacy. Crusan, Plakans, and Gebril (2016) developed a survey to examine the perspectives and background of writing teachers on assessment. Djoub (2017) developed a questionnaire for measuring the effectiveness of LAL on practices of assessment and Giraldo (2018) developed a questionnaire for LASs among teachers. The present study adopted the content of this recent questionnaire as it best fits the context of PY programs.

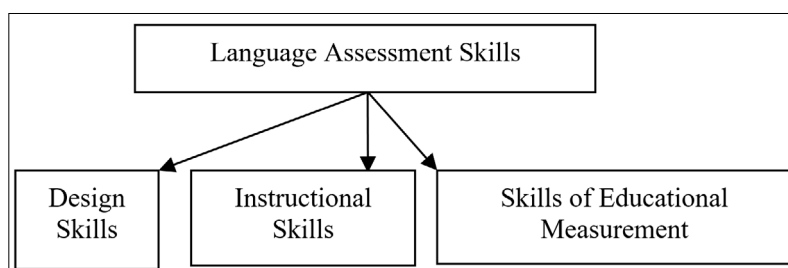
Methods

Approach

In language assessment, unique stakeholders need different levels of LASs. Language teachers need three sub-skills in language assessment. Giraldo (2018) proposed a list of LASs for language teachers. These skills are categorized as instructional skills, design skills, skills of educational measurement, and technological skills. The last two categories can be combined in one category where technological software is used for educational measurement. This list of skills coincides with Stabler-Havener’s (2018) groupings of LASs: skills in item-writing, skills in test construction, and skills in analysis. Accordingly, the present study classifies list of assessment skills as instructional skills, design skills, and skills of educational measurement (Figure 1).

Figure 1

Language Assessment Skills (LASs)



Source: (own elaboration)

Design skills refer to skills for writing test syllabuses that are based on learning outcomes (Taylor, 2009). These design skills also include true and false questions, multiple-choice questions, and so on (Giraldo, 2018). Design skills also involve the sub-skills of using rubrics and the skills for providing security for soft and hardcopies of tests. Writing exam proposals, which includes presenting, sequencing, and assisting in the achievement of learning objectives, is an integral part of design skills. Accordingly, faculty members should be able to design assessments that are “reliable, authentic, fair, ethical, practical, and interactive” (Fulcher 2012, p.116).

Instructional skills refer to the process of collecting formal and informal data for students’ language development. Formal means of data include portfolios and tests (Rea-Dickin, 2001). Informal data includes classroom observation. It also involves the skills of giving feedback and motivating students based on their results (Hill & McNamara, 2012). Instructional skills also subsume skills of reporting and interpreting results to various stakeholders such as students and parents (Giraldo, 2018). It is assumed that a firm foundation in instructional skills sharpens LASs among faculty members.

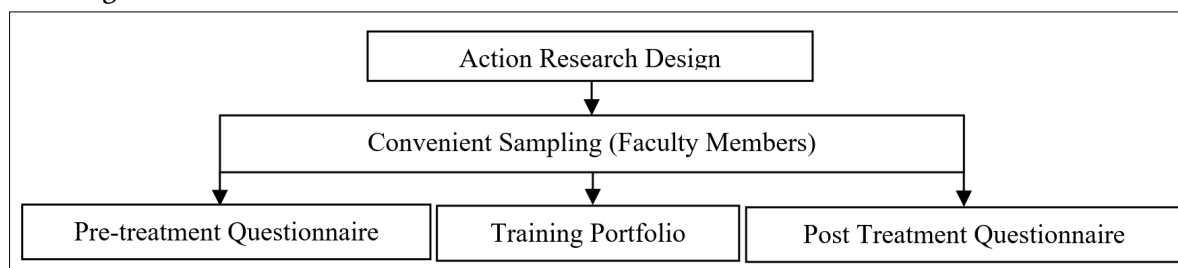
Skills of educational measurement include the ability to use statistical software such as Excel and SPSS. In so doing, faculty members will be able to calculate descriptive statistics such as means, curves, standard deviations, reliability, and correlations. Skills of educational measurement go beyond analysis into the interpretation of data showing students’ strengths and weaknesses (Giraldo, 2018). This view is supported by Davies (2008) and Inbar-Lourie (2012).

Several training approaches guide language assessment (Malone, 2017). Recently, Boyd and Donnarumma (2018) suggested a trainee-centered approach for LAL. In general, this approach deals with teachers not as individual novices, but as collective experts. This approach builds teachers’ engagement and confidence regarding assessment. Another feature of this approach is its flexibility for the issues of time and resources. It is a form of inductive learning that is suitable for workplace training. The trainer is a ‘regulator’ and the participants play the role of problem solvers. The current study uses this approach for its need-based training program.

Research Design

This study employed action research (Figure 2). First, a needs analysis questionnaire was administered to a convenient group of male faculty members because of the gender segregated system in Saudi universities. During the treatment, three workshops were conducted to address their needs in the form of a training program. After each workshop, a workshop evaluation form was distributed among the participants. At the end of the program, a post-treatment questionnaire was used.

Figure 2
Research Design



Source: (own elaboration)

Setting and Participants

Thirty-five male faculty members deliver six English courses across approximately 30 sections in the Department of English Language Skills in the PY program of Saudi University. These faculty members joined the department with different degrees (Ph.D. [n=8], Master’s [n=24] and BA [n=3], different teaching experiences (1-15 years), and various statuses of English language ability (EFL, ESL/NS). In terms of age, 45.2% of the

participants are between 30 and 39 years of age; 35.4% are over the age of forty; and about 20% of the participants are under the age of thirty (Table 1).

Table 1
Background Information of Participants

Age	Percent	Teaching Experience	Percent	Qualification	Percent
25-29	19.4%	1-5 years	25.8%	BA	9.7%
30-39	45.2%	6-10 years	29%	Master	71%
+40	35.4%	11+ years	45.2%	Ph.D.	19.3%

For teaching, 45.2% of the participants have more than ten years of experience. As their terminal degree, 71% of the participants hold a master's degree in English. The researchers informed the participants about the nature of the study and received their consent to use the data for research purposes with anonymity.

Faculty members have different overlapping roles and shared responsibility in harmonized assessment as classroom teachers, test writers, course coordinators, and/or exam committee members. All faculty members are considered classroom teachers. For testing, however, the department initially invited proposals from selected faculty members (test writers). They are provided with an assessment plan for which they must design questions based on intended learning outcomes. They then send the proposals to the course coordinators. The coordinators review the proposals and prepare one suggested exam draft, which is then sent to the exam committee. The exam committee reviews the items based on the characteristics of a good test and prepares valid and reliable tests. Necessary adjustments are made after multiple reviews. The exam paper is proofread again before it is printed.

The exam is administrated according to the schedule of the central examination committee. Exam papers are collected and distributed for evaluation, which is conducted by groups of 2-3 members. An exam paper goes through the various stages of evaluation, marking, rechecking, and teacher filtering. Marks are inserted into Excel spreadsheets for data analysis including question-wise analysis, learning outcomes, figures, means and standard deviations. Therefore, the target faculty members have overlapping roles in terms of designing and interpreting high-stakes assessment.

Instruments

Two instruments were used in this study: a questionnaire and a training portfolio. The pre-treatment questionnaire consisted of three parts: personal information, format and delivery method, and content of the questionnaire. Personal information included their name (optional), age, terminal degree, and teaching experience. Format and delivery method involve faculty members' preferences in terms of delivery format and trainer. The content of the questionnaire emphasized three main areas of LASs: design skills (12 items), instructional skills (12 items), and skills of educational measurement (8 items). A 5-point Likert scale was used; (from strongly disagree to strongly agree). The pre-treatment questionnaire was then distributed among the faculty members to identify their levels of LASs.

The data was analyzed using SPSS 20 software in which several analyses were employed: descriptive analysis, independent t-test, and paired t-test. Cronbach's Alpha was 0.974, which indicates that the questionnaire was reliable for internal consistency. After the needs analysis, the study addressed the reported needs in the form of a training program. The post-treatment questionnaire followed the content of the questionnaire used in the needs analysis. Cronbach's Alpha was 0.986.

The training portfolio was another instrument used in this study. It should be a systematic and organized tool based on participants' reflections and assessments (Khan & Begum, 2012). A training portfolio consists of observation sheets, training materials, and workshop evaluation form (Binet, Gavin, Carroll, & Arcaya, 2019). On the workshop evaluation form, the researchers asked the participants to point out their level of satisfaction and engagement with the workshops. A 1 to 5 scale was used where 1 showed the least level of satisfaction, and 5 the highest level of satisfaction. The evaluation forms were analyzed by calculating the degree of engagement under each scale. The observations sheets were analyzed using thematic analysis.

Procedure

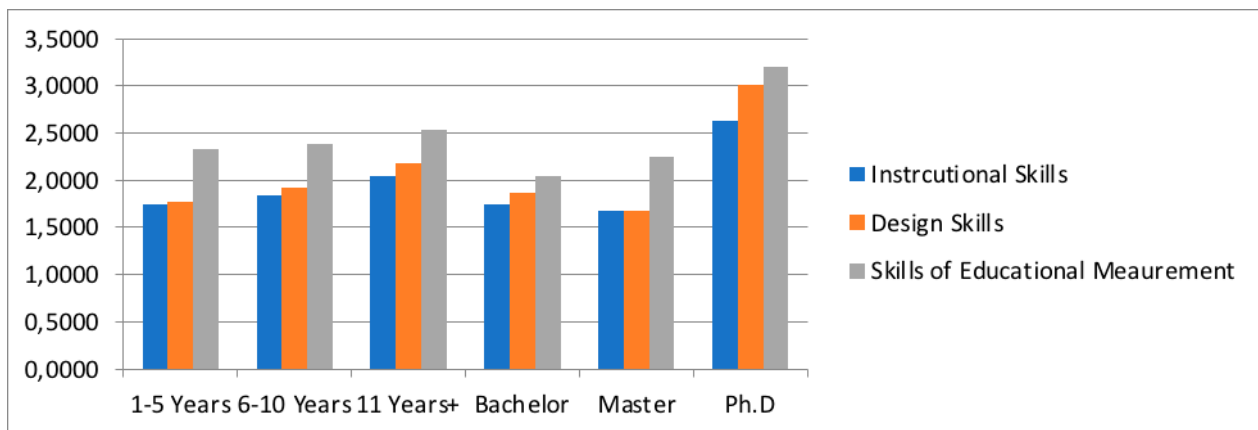
Needs Analysis

Thirty-one faculty members responded to the questionnaire. The two authors, Ph.D. holders, and two master's holders did not respond to the questionnaire. To collect objective data, the two authors avoided responding to the items on the questionnaire because their responses might affect the results of the study. The researchers analyzed the responses of the participants on the format of activities for LASs. As for the activities, the participants preferred workshops, group discussions, and seminars: 45%, 35%, and 20%; respectively. While 64.5% preferred conducting workshops at the workplace, 35.5% preferred online activities. As for preferred speakers, 32.5% of the participants wanted their colleagues to share their experiences. While 45% of the respondents preferred an expert from outside the institution, 22.5% preferred an expert from inside the institution.

The needs analysis shows low skill levels for educational measurement, design skills, and instructional skills (2.41, 1.97, 1.86); respectively. Then, the data analysis was conducted to see the relationships between LASs and teaching experience as well as qualifications (Figure 3).

Figure 3

LASs vs. Teaching Experience and Qualification (Needs Analysis)



It was found that teaching experience had no effect on instructional skills, design skills, and skills of educational measurement. All values were below the average except for skills of educational measurement where the participants have more than ten years of experience. Their value reached the average ($m=2.53$). This result can be attributed to assessment based on learning outcomes being a new experience for the participants who used to use content-based assessment. On the contrary, qualifications had a role in shaping the levels of LASs. Compared with other qualifications, Ph.D. holders ($n=6$) were found to be somewhat in need on skills of educational measurement ($m=3.2$), design skills ($m=3$), and instructional skills ($m=2.6$). Both holders of bachelor's degrees ($n=3$) and master's degrees ($n=22$) were still highly in need of training on LASs.

Accordingly, the researchers designed a training program. As most of the participants were master's degree holders, the researchers designed the training program to suit their need level and requested that all of the participants join the training program. Similarly, the researchers used workshops and group discussion activities. In addition, the activities were conducted at the workplace and online tutorials were shared among the participants. Although 45% of the participants preferred external experts, the researchers conducted the training program with an internal expert from the e-learning unit. It was beyond the researchers' limit to invite and support external experts to train the participants on LASs.

Training Program

In the context of professional development, the researchers as 'regulators' designed and conducted a training program consisting of six one-hour training sessions and two-and-a-half-hour online tutorials. The researchers reviewed the training materials between workshops and the reported needs to best enhance LASs among the participants. In addition, the researchers are qualified enough to facilitate this job since the first author is the

head of the examination committee and the second author's area of expertise are language assessment (Hazaea & Tayeb, 2018) and professional development (Hazaea, 2019). The participants, through the workshops, were trained (Ørngreen & Levinsen, 2017) on LAS areas as revealed by the needs analysis report. Each training session started with a theoretical orientation and then a group discussion. While two of the workshops were conducted by the two authors, one workshop was conducted by an expert from the e-learning unit.

To address the need for design skills, the first author conducted a training workshop entitled "Testing the test". It aimed to investigate question papers in terms of process and product. It involved the planning and preparation of exam proposals. It covered assessment methods and test construction. Using the revise engineering method (Walters, 2010), the participants were divided test-wise into four groups: Listening and Speaking, Reading, Writing, and Grammar. First, each group was given the submitted proposals for the previous final exam. They were then asked to evaluate the exam proposals and to find the strengths and weaknesses, and give suggestions for improving the exam proposals. After that, each group was given the final version of the same exam papers. They evaluated them in the same way they evaluated the exam proposals. Finally, each group compared the exam proposal with the final version of the exam papers.

To address the need for instructional skills, the second author conducted a training workshop entitled "Enhancing instructional skills for language assessment". The workshop first contextualized language assessment as high-stakes assessment. It then explained instructional skills for English courses. Because of time constraints and to cope with the participants' preferences, the second author shared *YouTube* links with the participants on three topics: How to use 1) Blackboard, 2) Google forms and 3) students' portfolios for language assessment. During group discussions, the participants were divided course-wise into four groups. They were given copies of the first midterm papers, course learning outcomes, covered syllabi, and a question paper review template. They were asked to review the exam papers, align learning outcomes, content, and exam papers, and to think about the strengths, weaknesses and suggestions for improvement. As an individual long-term task, the participants were asked to collect their students' scripts (formal data) and the participants' observations (informal data), and analyze them for students' language development and feedback. The participants were also asked to introduce students' portfolios for assessing writing and reading classes, and to use Blackboard for continuous assessment of their students.

To address the need for educational measurement skills, a training workshop entitled "Excel program as an educational measurement tool" was conducted in the computer lab by an expert from the e-learning unit. It aimed to apply Excel software to language assessment. To highlight the importance of Excel, the expert trained the participants on the types of data (texts, numbers and equations, formulas) and on how to create them in the program. They were also trained on how to protect the data inside the worksheet as well as on how to protect the Excel file with a password. The participants were also trained on how to generate means, standard deviations, and use of figures to report students' results, learning outcomes, and question-wise analysis.

The training program was intensively conducted during the second midterm week, a period where faculty members can be involved in an assessment environment. During this week, faculty members had a chance to participate in the workshops. After each workshop, a workshop evaluation form was distributed among the participants. Later, a post-questionnaire was distributed among the participants to evaluate the effectiveness of the treatment.

Results

Engagement in the Treatment

The analysis of the training portfolio shows that the participants were satisfied with the training program. They became more engaged and confident about test design and interpretation.

The analysis of the observation sheets shows that participants asked many questions during the theoretical part, especially during the sessions on instructional skills and design skills. These participants were curious about some of the assessment terms. When they were asked about 'language assessment literacy', it was observed that there were no responses, which indicates that the term seemed to be new to them. They indicated

that although they practiced some instructional skills such as reporting slow learners and good performance, they depended on direct methods for assessment data. They rarely used informal assessment data tools. In terms of design skills, some participants were surprised at the processes and the hidden work needed to assemble an exam paper before its final production. Furthermore, some participants asked for the training materials to be shared after the workshops.

During the group discussions in the workshops, the researchers observed that the participants were debating, which indicates overlapping roles such as classroom teachers, test writers, course coordinators, and members of the exam committee. The researchers also recorded the following observations during the training sessions or later when they were mentioned by the participants.

Next time, I am going to align learning outcomes, assessment plan in my exam proposal.
 It is a fruitful chance for developing test skills especially alignment of learning outcomes with questions.
 Today, I got new understanding to testing ... I hope to participate in similar workshops.
 It is a perfect chance for triggering the essence of test interpretation.
 I learnt from the discussion to improve the current situation to widen skills about testing and framing a balanced fair test to assess learners.
 The exam questions should be based on learning outcomes.
 I have to keep in mind Excel measurement when designing a test proposal.

The most challenging part for the participants was aligning learning outcomes, content, and the assessment plan. For test designing, the participants showed that they learnt how to construct good test structures. The participants also realized that they had different roles and shared responsibility in language assessment. They also realized that students' results are reported in several ways for different purposes. Students' results are reported to the competitive programs based on the highest scores where program of medicine, for example, selects the highest level of students. However, students' results are reported to other programs as 'pass' and 'fail'

The researchers also observed that the allocated time was not sufficient to address the needs for instructional skills and design skills. During the discussion sessions, the researchers had to stop the group work and collect data while the participants still needed more time to discuss the issues of the workshops. To overcome that challenge, the researchers shared the training materials with the participants. They also shared online tutorials for them to watch during their free time. During the workshop on measurement skills, the participants faced some technical problems with the Excel software including opening shared worksheets, and adding and calculating mathematic formulas.

The analysis of the workshop evaluation forms shows that the participants were satisfied and engaged with the new experience (Table 2).

Table 2
Satisfaction and Engagement with the Workshops

No.	Level of satisfaction and engagement	5	4	3	2	1
1.	I am satisfied with the topics, the tasks, and the structure of the workshop.	58%	32%	10%	0	0
2.	I have interacted and participated in the workshop.	45%	39%	16%	0	0
3.	The workshop is appropriate for me.	87%	13%	0	0	0
4.	I have developed new understandings of testing.	81%	19%	0	0	0
5.	I have improved in terms of testing practices as a result of the workshop.	58%	23%	13%	6%	0
6.	The skills practiced during the workshop can be applied in language testing.	94%	6%	0	0	0
7.	The skills practiced during the workshop can be transferred to my evaluation and marking situations.	87%	13%	0	0	0
8.	The information can be transferred to interpreting and reporting test results.	81%	19%	0	0	0

Source: (adapted from Boyd and Donnarumma, 2018 pp.123-125)

This table shows that the participants built their confidence in testing. Most of the participants reported that the information gained during the workshop could be applied in language testing. They also showed that the workshops were appropriate for them, and the skills practiced during the workshops could be transferred to their evaluation and marking situations.

Enhancing LASs

The participants improved their LASs. The independent t-tests of the pre-treatment and post-treatment (see appendix) showed that the participants increased their instructional skills, design skills, and skills of educational measurements to 4.25, 4.21, and 4.04; respectively. Similarly, the paired t-test showed significant differences between the pre-treatment and the post-treatment at a 95% confidence interval of the differences (Table 3).

Table 3
Paired Differences of the t-test

Pre-treatment – Posttreatment of LASs	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Dev.	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Instructional Skills	-2.38440	.31262	.09025	-2.58303	-2.18577	-26.421	31	.000
Design Skills	-2.26345	.41716	.12042	-2.52850	-1.99840	-18.796	31	.000
Skills of Educational Measurement	-1.62901	.42059	.14870	-1.98064	-1.27739	-10.955	31	.000
LASs	-2.15020	.48236	.08527	-2.32411	-1.97629	-25.216	31	.000

Table 3 shows the differences in the means and the statistical levels of significance of the pre-treatment and post-treatment of the LASs. Instructional skills and design skills scored very high means, 2.38 and 2.26, with a standard deviation of .312 and .417 respectively. Skills of educational measurement scored a high level, 1.62, with a standard deviation of .420. All in all, the paired test revealed the appearance of very high significances at the level of each LAS and the total LASs (.000).

The independent analysis of the post-treatment questionnaire The results of the treatment in the appendix shows that the participants improved instructional skills.

I can improve instruction based on assessment results and feedback. (57%)

I can provide feedback on students' assessment performance. (54%)

On the one hand, these two items indicated the highest level of improvement. On the other hand, the analysis showed that the participants need further training on using technology to assess students as in the following item *I can incorporate technologies in assessing students (35%)*. They also need to be trained to use alternative means such as the students' portfolios of language assessment as was indicated in the analysis of the following item. *I can utilize alternative means for assessment, for example, portfolios. (41%)*.

The independent analysis also showed the participants improved skills for designing tests (See appendix).

I can write selected-response items such as multiple-choice, true-false, and matching. (54%)

I can clearly define the language construct(s) a test will give information about. (53%)

These two items indicated that the participants highly improved in terms of item writing and language constructs. On the contrary, the participants were still in need of workshops for raters and for designing rubrics for alternative means of assessment as can be seen in the following least developed items.

I can design workshops for raters, whenever necessary. (22%)

I can design rubrics for alternative assessments such as portfolios and peer assessment (39%)

The participants also improved their skills in educational measurement (See appendix). As evidenced by the following two items, the participants can infer students' points of strengths and weaknesses based on the analysis of questions and learning outcomes. They also benefitted from the shared tutorial links about the use of Excel in language assessment.

I can infer students' strengths and weaknesses based on data. (45%)

I can use internet tutorials for particular language assessment needs. (43%)

However, the participants were still in need of critical analysis on external tests as evidenced by the following item analysis. *I can criticize external tests and their qualities based on their psychometric characteristics (22%)*.

This item reported the least developed measurement skill. Unexpectedly, the participants reported that they needed to use Excel for language assessment as it was evidenced by the analysis of the following item.

I can use software such as Excel for language assessment. (24%)

This result can be attributed to the nature of work in a PY program where the exam committee interprets and reports assessment data. The role of faculty members lies in inserting the data into the Excel sheets prepared by the exam committee. It seems that the participants need to know more about how Excel sheets are prepared and designed.

Discussion

This study aimed to scrutinize the impact of need-based training on LASs among faculty members in the PY program of Saudi University. First, it identified levels of LASs through a needs analysis. Then it designed a training program based on those reported needs; something to engage and motivate the participants regarding language assessment. Finally, it aimed to enhance LASs among the participants. This section revisits these three issues.

The nature of language assessment was a challenging endeavor for a PY program. Shifting the assessment to be based on learning outcomes had great effect on shaping the LASs among the participants. The needs analysis report showed that the participants lacked training on instructional skills, design skills, and skills of educational measurement. This need coincided with previous studies in the Saudi context. For example, Mansory (2016) found that EFL teachers at the Language Institute of Saudi University might be excluded from language assessment due to their need for developing LASs in technical areas, which necessitates technical preparation and training that can usually be expensive and time consuming. The study recommended long-term collaborative training and support that would greatly help in raising the assessment literacy among teachers. Similarly, Hakim (2015) reported poor practicing assessment techniques among EFL teachers at an English language institute in a Saudi university. The needs analysis was also in agreement with Hazaea and Tayeb (2018) who found that washback had the least effect on content assessment among the four factors of language teaching.

The present needs report is also supported by studies from European countries. On examining the training needs among language teachers in Europe, Vogt and Tsagari (2014) showed that teachers need training in language assessment. Because of inadequate training, teachers either learn about assessment at the workplace or assess their students with ready-made tests. Similarly, Boyd and Donnarumma (2018) found that EFL teachers need to be aware of the importance and impact of testing. They also need training on test design.

In Turkey, Öz and Atay (2017) suggested ongoing professional development sessions for EFL instructors to raise their awareness and to improve their performance in their own assessment. The same study found an imbalance between assessment literacy and classroom reflection. Hatipoğlu (2015) explored the role of assessment experience on EFL teachers' needs relevant to LAL.

In other countries, Lan and Fan (2019) found that EFL teachers in China were underdeveloped, especially in technical skills and language pedagogy. In Hong Kong, Lam (2015) concluded that potential research could investigate how in-service language teachers developed their LAL via workshops especially during the first few years of teaching. Sultana (2019) found that the insufficient academic background of assessment among Bangladeshi EFL teachers hindered their assessment skills.

Moving to another issue, the data analysis revealed that the participants engaged in the treatment. The training program moved the stagnant waters of language assessment for the participants. They participated actively in the workshops. These findings coincided with those reported by Boyd and Donnarumma (2018) who found that tutors were confident about testing after training sessions on principles of language assessment. In an assessment training program for EFL teachers in Egypt, the participants reported high levels of agreement regarding the program outcomes. The participants also stated that the workshops were very useful (Gebriel & Boraie, 2016). On the contrary, some disagreements and negotiations emerged during the group discussions

due to the overlapping roles of the participants. These findings coincide with Malone (2017), who reported certain challenges including disagreements between language testers and teachers about the topics to be covered in an exam paper.

After the treatment, the data analysis shows that the training program was effective in terms of enhancing LASs (2.15). As the paired t-test showed, instructional skills (2.38) and design skills (2.26) scored a very high level of significance. Skills on educational measurements (1.62) also scored a high level of significance. These findings coincided with the existing research. Walters (2010) found that “prior formal training in test specification writing (or lack of it) affected [Standard Reverse Engineering] processes and outputs” (p.337). Umer et al. (2018) found an overlap between course learning outcomes and question papers in an English language program at Saudi University, which can lead to surface-level learning among students. The findings also coincided with Xie and Lei (2019) who proposed instructional assessment strategies with the L2 writing instruction processes. Chen and Zhang (2017) also found that feedback from teachers played a vital role in EFL writing.

Yet, the participants still need further training on online assessment, assessing students’ portfolios, designing assessment workshops and rubrics, criticizing external tests, and programming Excel sheets for language assessment. These results suggest that training without practice is not effective. Some quizzes are sent online through the Blackboard, but online assessment where faculty members use Blackboard, for example, to generate questions, to give immediate feedback to students is not practiced in the PY program. That is why the participants feel that they need further training on online assessment. Similarly, it seems that the participants never use student portfolios in their assessment.

Guided by the LAL approach, the present study focused only on language teachers. On the contrary, students are still an important, yet neglected, stakeholder in language assessment. As teaching has moved into student-centered learning, student-centered assessment (Baranovskaya & Shaforostova, 2017) is still under-investigated among LAL theorists. For better language learning, recent research has shown the importance of involving students in their ‘alternative’ language assessment. For example, students can participate in their continuous assessment using student portfolios (Llarenas, 2019), peer assessment (Stognieva, 2015), self-assessment (Baranovskaya & Shaforostova, 2017), Olympiads (Bolshakova, 2015) and roundtable discussions (Rodomanchenko, 2017) to name a few.

Conclusions

This context-specific study reports the levels of LASs among EFL faculty members and the effectiveness of professional training on LASs for designing and interpreting tests. It first identified three levels of LASs, and then addressed those needs in the context of professional development. The needs analysis report showed that the participants need training in the three overlapping areas of instructional skills, design skills, and educational measurement skills. During treatment, the study found that the participants reported their confidence and satisfaction with language assessment. The analysis of the post-treatment questionnaire showed that the participants improved their instructional skills, design skills, and skills of educational measurements.

These improvements could have positive effect on the participants’ assessment practices. These findings imply constructive, valid, and fair assessments for PY students. These students are the indirect beneficiary of this experiment. Students’ voices also need to be heard in language assessment. In a student-centered environment, students can be involved in pre-test preparation. For example, once students produce their own quizzes, classroom teachers can better understand the way students think of the exam and can address their needs in language assessment accordingly. In the context of high-stakes tests, students always ask about the nature of the exam. Such questions can be answered in the context of learning outcomes. In other words, when students ask about the nature of the test, faculty members need to highlight the intended learning outcomes of the courses. However, further research is recommended on enhancing LASs among students.

If we were to repeat the training program, we would make several adjustments. First, we would include other stakeholders such as EFL schoolteachers, students, and the coordinators of the concerned programs that stipulate PY high scores before admission. Second, we would invite experts from outside the university to share in training faculty members. We would sort out the need for assessment skills for reading, vocabulary, writing, speaking, and listening. The allocated time for the training program was not enough to accommodate the busy schedules of faculty members. An intensive program on design and instructional skills could be better implemented during a semester break. A remedial training program could also be implemented to re-address the skills that still need further training and discussion. Other researchers could also employ a similar training program to their unique teaching contexts.

Acknowledgements

We would like to thank the Ministry of Education in Saudi Arabia and the Deanship of Scientific Research at Najran University for their financial funding and technical support for this project in the eighth research phase grant code No. NU/ SHED/16/109. Our gratitude extends to the editorial board and the anonymous reviewers of the manuscript for their constructive comments.

Conflict of interests

The authors declare that they have no conflict of interest.

References

- Baranovskaya, T., & Shaforostova, V. (2017). Assessment and evaluation techniques. *Journal of Language and Education*, 3(2), 30-38. <https://doi.org/10.17323/2411-7390-2017-3-2-30-38>
- Binet, A., Gavin, V., Carroll, L., & Arcaya, M. (2019). Designing and facilitating collaborative research design and data analysis workshops: Lessons learned in the healthy neighborhoods study. *International journal of environmental research and public health*, 16(3), 324-339. <http://dx.doi.org/10.3390/ijerph16030324>
- Bolshakova, E. (2015). Olympiad in the English language as a form of alternative language assessment. *Journal of Language and Education*, 1(2), 6-12. <https://doi.org/10.17323/2411-7390-2015-1-2-6-12>
- Boyd, E., & Donnarumma, D. (2018). Assessment literacy for teachers: A pilot study investigating the challenges, benefits and impact of assessment literacy training. In D. Xerri & P. V. Briffa (Eds.), *Teacher involvement in high-stakes language testing* (pp. 105-126). Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-77177-9_7
- Brindley, G. (2001). Language assessment and professional development. In C. Elder, A. Brown, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davies* (pp. 137-143). Cambridge University Press.
- Chen, D. & Zhang, L. (2017). Formative assessment of academic English writing for Chinese EFL learners. *TESOL International Journal*, 12(2), 47-64
- Crusan, D., Plakans, L., & Gebril, A. (2016). Writing assessment literacy: Surveying second language teachers' knowledge, beliefs, and practices. *Assessing Writing*, 28, 43-56. <http://dx.doi.org/10.1016/j.asw.2016.03.001>
- Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, 25(3), 327-347. <http://dx.doi.org/10.1177/0265532208090156>
- Djoub, Z. (2017). Assessment literacy: Beyond teacher practice. In R. Al-Mahrooqi, C. Coombe, F. Al-Maamari & V. Thakur (Eds.), *Revisiting EFL assessment: Second language learning and teaching* (pp. 9-27). Springer. http://dx.doi.org/10.1007/978-3-319-32601-6_2
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly*, 9(2), 113-132. <http://dx.doi.org/10.1080/15434303.2011.642041>
- Galichkina, E. (2016). Developing teacher-trainees' assessment awareness in the EFL classroom through project-based learning activity. *Journal of Language and Education*, 2(3), 61-70. <http://dx.doi.org/10.17323/2411-7390-2016-2-3-61-70>

- Gebril, A., & Boraie, D. (2016). Assessment literacy training for English language educators in Egypt. In V. Aryadoust & J. Fox (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp. 416-437). Cambridge scholars publishing.
- Giraldo, F. (2018). Language assessment literacy: Implications for language teachers. *Profile Issues in Teachers' Professional Development*, 20(1), 179-195. <http://dx.doi.org/10.15446/profile.v20n1.62089>
- Hakim, B. (2015). English language teachers' ideology of ELT assessment literacy. *International Journal of Education & Literacy Studies*, 3(4), 42-48. <http://dx.doi.org/10.7575/aiac.ijels.v.3n.4p.42>
- 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.
- Hazaea, A. (2019). The needs on professional development of English language faculty members at Saudi University. *International Journal of Educational Researchers*, 10(1), 1-14
- Hazaea, A. N., & Tayeb, Y. A. (2018). Washback effect of LOBELA on EFL teaching at preparatory year of Najran University. *International Journal of Educational Investigations*, 5(3), 1-14
- Herrera M., L., Macías, V., & Fernando, D. (2015). A call for language assessment literacy in the education and development of teachers of English as a foreign language. *Colombian Applied Linguistics Journal*, 17(2), 302-312. <https://doi.org/10.14483/udistrital.jour.calj.2015.2.a09>
- Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language testing*, 29(3), 395-420. <http://dx.doi.org/10.1177/0265532211428317>
- Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25(3), 385-402. <http://dx.doi.org/10.1177/0265532208090158>
- Inbar-Lourie, O. (2012). Language assessment literacy. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-9). John Wiley & Sons. <http://dx.doi.org/10.1002/9781405198431.wbeal0605>
- Khan, B., & Begum, S. (2012). Portfolio: A professional development and learning tool for teachers. *International Journal of Social Science and Education*, 2(2), 363-377.
- Lam, R. (2015). Language assessment training in Hong Kong: Implications for language assessment literacy. *Language Testing*, 32(2), 169-197. <http://dx.doi.org/10.1177/0265532214554321>
- Lan, C., & Fan, S. (2019). Developing classroom-based language assessment literacy for in-service EFL teachers: The gaps. *Studies in Educational Evaluation*, 61, 112-122 <https://doi.org/10.1016/j.stueduc.2019.1003.1003>.
- Language Assessment Literacy Symposium. (2016). Enhancing language assessment literacy: Sharing, broadening, innovating. *Language Assessment Literacy*. URL <http://wp.lancs.ac.uk/ltrg/files/2015/12/Language-Assessment-Literacy-Symposium-2016-full-programme.pdf>
- Llarenas, A.T. (2019). Vignettes of experiences in portfolio assessment. *Asian EFL Journal*, 21(2), 196-223
- Malone, M. E. (2017). Training in language assessment. In E. Shohamy, L. G. & S. May (Eds.), *Language Testing and Assessment, Encyclopedia of Language and Education* (vol. 3, pp. 225-239). Springer. http://dx.doi.org/10.1007/978-0-387-30424-3_178
- Mansory, M. (2016). *EFL teachers' beliefs and attitudes towards English language assessment in a Saudi university's English language institute* [Unpublished doctoral dissertation]. University of Exeter. URL <http://hdl.handle.net/10871/25765>
- Ørngreen, R., & Levinsen, K. (2017). Workshops as a research methodology. *Electronic Journal of E-learning*, 15(1), 70-81.
- Öz, S., & Atay, D. (2017). Turkish EFL instructors' in-class language assessment literacy: Perceptions and practices. *ELT Research Journal*, 6(1), 25-44.
- Pill, J., & Harding, L. (2013). Defining the language assessment literacy gap: Evidence from a parliamentary inquiry. *Language Testing*, 30(3), 381-402. <http://dx.doi.org/10.1177/0265532213480337>
- Rea-Dickins, P. (2001). Mirror, mirror on the wall: Identifying processes of classroom assessment. *Language Testing*, 18(4), 429-462. <http://dx.doi.org/10.1177/026553220101800407>
- Rodomanchenko, A. (2017). Roundtable discussion in language teaching: Assessing subject knowledge and language skills. *Journal of Language and Education*, 3(4), 44-51. <https://doi.org/10.17323/2411-7390-2017-3-4-44-51>
- Stabler-Havener, M. L. (2018). Defining, conceptualizing, problematizing, and assessing language teacher assessment literacy. *Working Papers in Applied Linguistics and TESOL*, 18(1), 1-22.
- Stognieva, O. (2015). Implementing peer assessment in a Russian university ESP classroom. *Journal of Language and Education*, 1(4), 63-73. <https://doi.org/10.17323/2411-7390-2015-1-4-63-73>
- Sultana, N. (2019). Language assessment literacy: an uncharted area for the English language teachers in Bangladesh. *Language Testing in Asia*, 9(1), 1-14. <http://dx.doi.org/10.1186/s40468-019-0077-8>
- Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21-36. <http://dx.doi.org/10.1017/S0267350X09000000>

org/10.1017/S0267190509090035

- Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing*, 30(3), 403-412. <http://dx.doi.org/10.1177/0265532213480338>
- Umer, M., Zakaria, M. H., & Alshara, M. A. (2018). Investigating Saudi University EFL teachers' assessment literacy: Theory and practice. *International Journal of English Linguistics*, 8(3), 345-356 <http://dx.doi.org/10.5539/ijel.v8n3p345>
- Vogt, K., & Tsagari, D. (2014). Assessment literacy of foreign language teachers: Findings of a European study. *Language Assessment Quarterly*, 11(4), 374-402. <http://dx.doi.org/10.1080/15434303.2014.960046>
- Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering. *Language Assessment Quarterly*, 7(4), 317-342. <http://dx.doi.org/10.1080/15434303.2010.516042>
- Xie, Q., & Lei, Y. (2019) formative assessment in primary English writing classes: A case study from Hong Kong. *Asian EFL Journal*, 23(5), 55-95

Appendix

Language Assessment Skills

Statement	Pre-treatment		Post Treatment		Improv.
	Means	Std. Dev	Means	Std. Dev	Percent
Instructional Skills	1.868283	0.939779	4.252683	0.797118	47%
I can improve instruction based on assessment results and feedback.	1.5161	.81121	4.3548	.75491	57%
I can provide feedback on students' assessment performance.	1.5484	.76762	4.2581	.77321	54%
I can provide motivating assessment experiences, giving encouraging feedback.	1.5806	.76482	4.2581	.81518	54%
I can collect formal data (e.g. through tests) on students' language development.	1.6774	.79108	4.3226	.70176	53%
I can plan, implement, monitor, record, and report student language development.	1.8710	.84624	4.3548	.75491	50%
I can align learning outcomes, instruction, and assessment.	1.8710	.92166	4.3226	.74776	49%
I can collect informal data (e.g. while observing in class) on students' language development.	1.9677	.94812	4.2903	.82436	46%
I can use language assessment methods appropriately.	1.8387	1.00322	4.1613	.82044	46%
I can use multiple methods of assessment to make decisions based on substantive information.	2.0323	1.01600	4.2581	.92979	45%
I can communicate test results to a variety of audiences: students, department, and deanship.	2.0000	1.09545	4.1290	.84624	43%
I can utilize alternative means for assessment, for example, portfolios.	2.0968	1.19317	4.1290	.76341	41%
I can incorporate technologies for assessing students.	2.4194	1.11876	4.1935	.83344	35%
Design Skills	1.970664	0.990725	4.212367	0.820873	45%
I can write selected-response items such as multiple-choice, true-false, and matching.	1.7742	.95602	4.4516	.72290	54%
I can clearly define the language construct(s) a test will give information about.	1.6774	.79108	4.3226	.74776	53%
I can clearly identify and state the purpose for language assessment.	1.7097	.78288	4.2258	.84497	50%
I can design assessments that are valid not only in terms of course contents but also course tasks.	1.7419	.81518	4.1935	.79244	49%
I can improve test items after item analysis, focusing on items that are either too difficult, too easy, or unclear.	1.8065	1.01388	4.2581	.89322	49%
I can design assessments that are reliable, authentic, fair, ethical, practical, and interactive.	1.8710	1.02443	4.2258	.84497	47%
I can design constructed-response items (for speaking and writing), along with rubrics for assessments.	1.8710	.95715	4.1613	.82044	46%
I can construct test specification to design parallel forms of tests.	1.9032	1.01176	4.1613	.68784	45%
I can write test syllabuses to inform test users of test formats, where applicable.	1.9677	1.07963	4.2258	.84497	45%
I can provide security to ensure that unwanted access to tests is deterred.	1.9032	1.07563	4.0968	.94357	44%
I can design rubrics for alternative assessments such as portfolios and peer assessments.	2.1935	1.07763	4.1613	.77875	39%
I can design workshops for raters, whenever necessary.	2.9677	1.30343	4.0645	.92864	22%
Skills of Educational Measurement	2.415338	1.155141	4.04435	0.949653	33%
I can infer students' strengths and weaknesses based on data.	2.0000	.89443	4.2258	.88354	45%
I can use internet tutorials for language assessment needs.	2.1613	1.15749	4.2903	.90161	43%
I can interpret data related to test design such as item difficulty and item discrimination.	2.2903	1.21638	4.1290	.88476	37%
I can calculate reliability and validity indices by using appropriate methods such as Excel spreadsheets.	2.2581	1.18231	4.0000	.89443	35%
I can calculate descriptive statistics such as means and curves, reliability, and correlations.	2.4839	1.15097	3.8710	1.14723	28%
I can investigate facility and discrimination indices statistically.	2.5484	1.26065	3.9677	.87498	28%
I can use software such as Excel for language assessment.	2.6452	1.22606	3.8387	1.09839	24%
I can criticize external tests and their qualities based on their psychometric characteristics.	2.9355	1.15284	4.0323	.91228	22%