مجلة العلوم التربوية

The Perspectives of Saudi EFL Teachers on Practical Criteria of Test Construction

Hayat Rasheed Hamzah Alamri
College of Education, Taibah University, PO box 42374,
Madinah 4445, Saudi Arabia

The Perspectives of Saudi EFL Teachers on Practical Criteria of Test Construction

Hayat Rasheed Hamzah Alamri
College of Education, Taibah University, PO box 42374, Madinah 4445, Saudi Arabia
* E-mail of the corresponding author: hramri@taibahu.edu.sa

Abstract

The current descriptive quantitative study investigates the perspectives of Saudi EFL teachers on test construction criteria. The research also aims to understand their gender-specific views on test construction criteria. Therefore, the researchers developed a 32-5-point Likert scale questionnaire to measure teachers' perspectives. The sample includes 227 teachers, 83 males, and 144 females, during the 2021 and 2022 academic years. Results indicated that the test construction process is crucial for EFL teaching, learning, and assessment. EFL teachers generally have good test design and development knowledge, preferring objective items over subjective ones. They highly value the practical criteria of test construction and the essential role of all seven dimensions in the process. The Mann-Whitney U-test revealed no significant gender differences among EFL male and female teachers in the seven dimensions perspectives. The study concludes with recommendations for in-service EFL teachers to receive more comprehensive training and attend conferences, workshops, and seminars to enhance their test construction abilities.

Keywords: Assessment, test construction, criteria, perspectives, Saudi EFL teachers

المستخلص:

تبحث الدراسة الكمية الوصفية الحالية في وجهات نظر معلمي اللغة الإنجليزية كلغة أجنبية سعوديين حول معايير بناء الاختبار. ويهدف البحث أيضًا إلى فهم وجهات نظرهم الخاصة بالجنس حول معايير بناء الاختبار. ولذلك، قام الباحثون بتطوير استبيان بمقياس ليكرت مكون من ٣٢–٥ نقاط لقياس وجهات نظر المعلمين. وتضم العينة ٢٢٧ معلماً، ٨٣ ذكراً، و١٤٤ أنثى، خلال العامين الدراسيين ٢٠٢١ و٢٠٢٢. أشارت النتائج إلى أن عملية بناء الاختبار أمر بالغ الأهمية لتدريس اللغة الإنجليزية كلغة أجنبية وتعلمها وتقييمها. يتمتع معلمو اللغة الإنجليزية كلغة أجنبية بشكل عام بمعرفة جيدة بتصميم الاختبارات وتطويرها، ويفضلون العناصر الموضوعية على العناصر الذاتية. إنهم يقدرون بشدة المعايير العملية لبناء الاختبار والدور الأساسي لجميع الأبعاد السبعة في العملية. كشف اختبار مان ويتني U عن عدم وجود فروق ذات دلالة إحصائية بين الجنسين بين معلمي ومدرسي اللغة الإنجليزية كلغة أجنبية في منظور الأبعاد السبعة. وتختتم الدراسة بتوصيات لمعلمي اللغة الإنجليزية كلغة أجنبية أثناء الخدمة لتلقي تدريب أكثر شمولاً وحضور المؤتمرات وورش العمل والندوات لتعزيز قدراتهم في بناء الاختبارات.

الكلمات المفتاحية: التقييم، بناء الاختبار، المعايير، وجهات النظر، معلمي اللغة الإنجليزية كلغة أجنبية سعوديين

1. Introduction

Assessment is crucial to teaching and learning, indicating students' progress to teachers and institutions. Farangi and Rashidi (2022) argued that despite the importance of the assessment concept, it has been overlooked for years. Assessment gives students a clear vision of their final assessment, enabling them to plan and predict future exams (Abbas, 1994). Assessment is considered a mission that teachers undertake to collect data on teaching outcomes and use this information for further improvement (Ahmed, 2018). Moreover, good assessment provides valuable learner information that impacts curriculum and pedagogy (Alfallaj & Al-Ahdal, 2017). Dollah and Atmowardoyo (2022) defined assessment as "the process by which information is obtained relative to some known goal or target" (p. 11). Assessment data is vital for evaluating educational interventions, determining the appropriateness of teaching methods, syllabi, and teacher education, and serves as the basis for decision-making to exclude ineffective interventions or suggest new ones after each teaching phase (El-Hassan & Ahmed, 2023).

On the other hand, Arslan and Üçok-Atasoy (2020) pointed out that assessment is essential in teaching English as a Foreign Language (EFL) programs, as it mirrors teaching practices. It provides constant feedback on teaching effectiveness. It's fundamental to assess classroom practices to ensure they align with expected performance outcomes. More recently, Swaie (2023) highlighted the importance of assessment in language classrooms and the need for teacher education programs to address the lack of training and development in assessing students' language skills and called for professional development programs for pre-service and in-service teachers.

Abbas (1994) emphasizes the importance of evaluating the EFL assessment process to ensure it meets efficient criteria and provides reliable data for practical evaluation, decisions, and judgments about teaching practices. Abbas also emphasizes the integration of test construction and EFL teaching as a constraint or determiner.

Dollah and Atmowardoyo (2022, p. 11) clarified that a test is a particular form of assessment made under contrived circumstances, mainly so that it can be carried out. In other words, all tests are assessments, but not all assessments are tests. Moreover, Ahmed (2018) referred to tests as "the most common assessment tool in the educational systems in the world. Whether students like them or not, this is a fact, and they have to deal with them" (p.

51). Ahmed (2018) argued that testing in school is usually done for assessment to give students grades; however, testing offers valuable benefits such as aiding decision-making, identifying learning gaps, enhancing learning, facilitating knowledge transfer, providing feedback to teachers, and encouraging frequent testing to motivate students to study and improve, ultimately enhancing the learning process.

Therefore, Ali et al. (2019) claimed that teachers must possess strong test construction skills to avoid false assessments of students' outcomes. Farangi and Rashidi (2022) also asserted that teachers play a crucial role in educational systems, influencing assessment information for learning purposes. They construct, implement, and adjust tests for learners, diagnose learning difficulties, study curriculum impacts, and assess teachers' strengths and weaknesses. According to Evers and Walberg (2005), tests serve constructive purposes, such as diagnosing students' learning difficulties, enhancing teaching methods, and assessing teachers' strengths and weaknesses. Kustati and Zurniati (2019) emphasize the interconnection of test, teaching, and learning processes, emphasizing the need for teachers to understand this concept to use tests for student achievement effectively.

Numerous studies have examined the competencies required by EFL teachers to construct classroom-based tests. For example, Ivanova and Terzieva (2016) suggest using Bloom's taxonomy to help teachers create tests that promote the English language and higher-level cognitive thinking skills, covering six cognitive levels: knowledge, comprehension, application, analysis, synthesis, and evaluation.

Moreover, Ali et al. (2019) highlighted common issues in EFL contexts, including inadequate knowledge and training for EFL teachers in test-taking and grading writing scripts, practical constraints, administrative hurdles, and insufficient training in designing and implementing language tests. Belarbi and Bensafa (2020) suggested that EFL tests and exams should be designed with clear, concise questions to allow students to answer without wasting time on irrelevant questions. Similarly, Imsa-ard (2020) suggested that language tests should incorporate listening and speaking skills to prevent reliance on rote learning and memorization. The study by Işık, (2020) revealed students' dissatisfaction with traditional assessment practices, highlighting low school assessment quality and the need for comprehensive teacher training to improve assessment literacy among English teachers, as

conventional approaches focused on formal English properties were common. Likewise, Ramírez (2020) highlighted the disparity between academic expectations and institutional administration and policies, highlighting the lack of training and literacy in language testing.

Hirai et al. (2022) analyzed the English Critical Thinking Test's validity, revealing challenges in creating test items measuring higher-order critical thinking skills, such as internal consistency and separation of inference and analysis sections. Farangi and Rashidi (2022) analyzed Iranian EFL teachers' assessment conceptions and self-efficacy. Results showed teachers view assessment as a tool for measuring student learning, modifying teaching practices, and indicating school performance. Proficiency in questioning, assessing learning, and providing alternative explanations predicted students' engagement. On the other hand, Saher et al. (2022) suggest the need for institution-in-service courses for teachers to learn how to construct traditional tests, as they are their preferred assessment method.

Moreover, Sun (2022) identifies four factors causing unsatisfactory language assessment literacy (LAL) levels among in-service EFL teachers in China: examination-oriented environment, lack of policy communication, teachers' personal beliefs, and insufficient training.

More recently, Swaie (2023) examined classroom-based assessment by EFL teachers in Jordan's secondary schools and found that some factors influence teachers' choice of assessment methods, including the National Exam, students' proficiency level, and their assessment knowledge.

2. Statement of the problem

EFL assessment patterns have significantly evolved since English was introduced as a compulsory foreign language because the assessment process is the only means for obtaining data necessary to evaluate components of the educational process, and its absence or lack of punctuality would negatively affect the planning of the EFL teaching program. Generally, several studies (Al-Seghayer, 2014; Alghamdi, 2016; Alqahtani, 2019; Altheyab, 2023) revealed that Saudi EFL teacher training programs are criticized for being inadequate and non-systematic. Many graduates lack teaching experience and proper educational preparation, leading to undesirable outcomes. The programs often neglect current teaching methods, assessments, and students' development. Additionally, there is a lack of exposure to authentic English-

speaking teachers, as many instructors come from non-native Arabic-speaking countries.

Moreover, over 25 years of experience teaching and supervising EFL teachers in Saudi schools reveal that teachers often face problems preparing test questions for English language courses. Thus, the researcher analyzed 57 test papers prepared by EFL teachers. The analysis process results indicated several errors that affected the test quality. The mistakes included language errors, question type, content coverage, unclear directions, confusing questions, multiple-choice option lengths**,** and student-level unsuitability. Besides, it seemed that teachers lacked interest in writing general test instructions, which can cause ambiguity and affect students' test readability and performance abilities during the assessment process. Addressing these issues can improve the quality of EFL teachers in test construction competence.

Moreover, revisiting the previous research (Alfallaj & Al-Ahdal, 2017; Ali et al., 2019; Alrzini et al., 2022; Belarbi & Bensafa, 2020; Kustati & Zurniati, 2019) reveals that very few studies have been conducted on designing EFL test questions. Therefore, the current study aims to explore the perspectives of Saudi EFL teachers on the practical criteria for test construction and evaluate its effectiveness in EFL teaching settings. Besides, it tends to investigate the views of Saudi EFL teachers on the requirements for test construction based on their gender.

3. Research questions

1. How do Saudi EFL teachers perceive their knowledge about constructing language tests?
2. What test question types serve Saudi EFL teachers' language learning goals?
3. What are the perspectives of Saudi EFL teachers on the practical criteria for test construction?
4. What are the differences between male and female Saudi EFL teachers in their perspectives on the practical criteria for test construction?

4. Research hypothesis

H1. There are no statistically significant differences at 0.05 between EFL male and female Saudi EFL teachers' perspectives on the practical criteria for test construction.

5. Research significance

1. The study provides a comprehensive understanding of the practical criteria for test construction, enabling EFL teachers to assess their test design techniques.
2. This study explores EFL teachers' shortcomings in practical criteria for test construction.
3. It provides EFL teachers with feedback about their effectiveness as test designers.
4. It develops a clear vision in applying practical criteria for constructing classroom tests.

6. Review of literature

Assessing learners' outcomes is complex due to teachers' differing perspectives on the process. Teachers evaluate their daily classroom actions, which is crucial for teaching and learning development. However, it's now acknowledged that there are multiple ways to gather information about student learning, and testing is only one part of assessment.

Assessment

Assessment can impact students and teachers, serving various functions such as evaluating individuals and programs, holding groups accountable for specific outcomes, informing instruction, and determining access to programs or privileges (Kustati & Zurniati, 2019; Q. Sun & Bin-Sihes, 2020; Xu & Brown, 2017).

"Assessment refers to any process involved in evaluating performance. Assessment is a professional judgment process involving various activities such as writing questions, designing classroom assessments, creating rubrics, scoring student work, arriving at grades, and interpreting standardized test scores" (Caldwell, 2008, p. 26). According to Alokozaya (2022), assessment is categorized into traditional and alternative. Alternative assessment is used in specific contexts to evaluate students' performance and knowledge. It provides more accessible information to interpret and understand, unlike traditional tests, which are single-occasion, focus solely on scores, and are unidimensional, such as timed exercises in multiple-choice or short-answer forms. However, student learning was measured solely through testing in traditional school settings. Dikli (2003) highlighted that testing is formal and provides specific procedures for administering and scoring. On the other hand, assessment is based on student knowledge and

abilities and involves various methods for collecting information at different times and contexts.

Caldwell (2008) also indicated that assessment is formal and informal. Formal assessment uses commercially produced instruments with specific administration and scoring procedures. In contrast, informal assessment measures are flexible, allowing educators to modify and adapt designs to particular student needs or classroom situations. Informal assessment is viewed as more authentic by educators as it closely resembles the actual activities students perform in schools.

Caldwell (2008), Farhady and Selcuk (2022), Handrayani (2022), and Hidri (2020) also pointed out that formative and summative are two types of assessment, with formative focusing on instruction and the teacher taking responsibility for student learning, and summative concentrating on comprehension and learning, respectively. They added that teachers enhance classroom learning through various teaching and assessment methods. Despite concerns about teacher-based assessments, classroom-based assessment, either summative or formative, offers opportunities to improve learning. Formative assessment evaluates students' learning activities, focusing on comprehension and performance. It provides feedback and helps teachers select suitable teaching methods, materials, and academic support, ensuring students' progress is accurately measured. On the other hand, **s**ummative assessment measures students' progress in language learning through midterm tests and final exams, helping teachers determine students' capabilities and materials for the following lessons, ensuring effective learning and comprehension. Summative assessments provide accountability but may not provide detailed information about individual student progress.

Caldwell (2008, p. 32) outlines guidelines for successful classroom assessment, including clearly defining the purpose, linking assessment activities to classroom objectives and instruction, explaining the assessment to students, using multiple methods and experiences to evaluate performance, and keeping assessment simple. These guidelines aim to ensure practical and effective classroom evaluation.

Furthermore, Xu and Brown (2017) identified seven competency domains for teachers to be skilled in assessment: choosing appropriate assessment methods, developing them, administering and interpreting results, using assessment results for decision-making, valid pupil grading procedures,

communicating results to stakeholders, and recognizing unethical methods and inappropriate uses of assessment information.

Testing

Regarding tests, Bachman (2000) defined a test as "a procedure used to establish the quality, performance, or reliability of something, particularly before widespread use" (p. 2). Algarabel and Dasí (2001) define a test as "a tool driven by an interest in measuring mental abilities. Their conceptualization influenced the technology used in test construction, which has been applied for a long time" (p.43). Likewise, Ahmed (2018) recently defined tests as "the most common assessment tool in the educational systems in the world. Whether students like them or not, this is a fact, and they have to deal with them" (p. 51).

Teachers use test-based classroom assessments to evaluate student learning and scaffold comprehension, categorizing questions into literal, text, and application inference stages, covering knowledge, comprehension, application, analysis, synthesis, and evaluation (Caldwell, 2008). Clay (2001) categorized test items into objective and subjective. Objective items include multiple-choice, true-false matching, and completion, while the subjective items include short-answer, extended-response, problem-solving, and performance test items.

Testing has several benefits for schools, teachers, and students. Ahmed (2018) and Algarabel and Dasí (2001) emphasize that testing aids decision-making, identifies learning gaps, helps students learn for the next stage, develops knowledge transfer, provides teachers feedback, and inspires. Frequent testing also encourages students to study and improve, ultimately contributing to the overall quality of education.

On the other hand, Sun and Bin-Sihes (2020) found that many language teachers struggle with test construction, interpretation, and evaluation due to time and resource constraints. The complexity of writing test items, such as multiple-choice, true/false, matching, and reading comprehension, makes teacher-produced tests less relevant. Factors contributing to this include schools using test papers from other institutions, lack of research, training, and heavy workload. Additionally, the abundance of disciplined teaching materials and online resources leads to a low teacher-centered approach and more severe academic performance.

Moreover, Kopriva (2008) highlighted that EFL learners may have ancillary abilities that can distort results due to less ideal measurement methods, affecting the measurement of targeted abilities. In response, Belarbi and Bensafa (2020) suggested that EFL tests should be designed with clear, concise questions to avoid irrelevant questions and focus on preparing competent performers, critical thinkers, and efficient learners rather than solely focusing on high scores and grading without communicative competence.

Caldwell (2008) recommends following guidelines for assessing comprehension using questions. These include distinguishing between literal and inferential questions, selecting response and constructed responses, aligning questions with instruction objectives, distinguishing between formative and summative assessment, noting that incorrect answers may indicate language or format issues, establishing assessment guidelines for open-ended questions, considering look-backs, and organizing user-friendly questions.

Henning (2012) identified common testing mistakes for EFL teachers, including general examination characteristics, insufficient number of items, redundancy of test types, lack of confidence measures, negative washback, trick questions, redundant wording, divergence cues, and option number. Test-value concerns include mixed content, wrong medium, shared knowledge, syllabus mismatch, and content matching. Administrative and scoring issues include a lack of cheating controls, inadequate instructions, administrative inequities, lack of piloting, and subjectivity of scoring.

To conclude, assessment and testing are distinct concepts. Assessment focuses on understanding a student's current knowledge, and testing is a single-time, one-dimensional exercise used to measure student learning in schools. Assessment involves multiple methods for collecting information at different times and contexts, while testing is seen as only one part of assessment. Both methods are essential for understanding student learning.

7. Previous studies

Limited research investigating studies on the EFL teachers' practices of test construction competence may offer insights into understanding how tests as assessment tools are implemented in EFL contexts. For example, Lemmetti (2015) emphasized the importance of validity, reliability, and washback in constructing and using a test. The study found that most

classroom tests are unreliable or valid due to teachers' inability to construct proper tests with these features.

Using a validated questionnaire and scenario-based test, Farhady and Tavassoli (2018) developed a test to measure EFL teachers' language assessment knowledge (LAK). Results showed low LAK levels among teachers, with potential relationships between factors like gender, field, education, and experience, emphasizing LAK's importance in EFL teaching.

Kustati and Zurniati (2019) explored EFL teachers' challenges in creating an English Mid-Semester test at SMPN 12 Padang. Data was collected through interviews and document analysis. The test is based on principles' association, with teachers and non-teachers working together. The study recommends giving the test to schoolteachers and committees who are well-versed in students' abilities and comprehension.

Trisanti (2019) explored the impact of assessment literacy (AL) on EFL teachers in Central Java, Indonesia. Over ten teachers were observed, revealing a lack of understanding in designing classroom assessments, setting test specifications, and devising appropriate test types. The findings suggest that teachers must improve their AL to ensure high-quality assessments.

The study by Abahussain et al. (2020) at Majmaah University, Saudi Arabia, analyzed 250 final exam questions from two semesters. It found balanced, valid, and reliable questions with four major themes: standardization of questions, Bloom's taxonomy efficacy, assessment/evaluation, and alignment with learning outcomes. The study recommends exams for maximum course coverage, question variety, clear instructions, validity, and alignment with learning outcomes.

The study by Arslan and Üçok-Atasoy (2020) found inconsistencies between English language teaching policy and assessment practices in middle schools in Turkey. The study involved 152 EFL teachers and used quantitative and qualitative methods. The results revealed that traditional tests were designed based on language structures and vocabulary rather than communicative competence or language skills.

Belarbi and Bensafa (2020) examined the effectiveness of Algerian EFL baccalaureate exam papers, focusing on their coverage of lower and higher-order thinking skills. The analysis revealed that the exam does not assess students' communicative abilities or higher-order thinking skills,

suggesting the need for practical EFL exams to enhance students' language competencies.

Sun and Bin-Sihes (2020) studied EFL secondary school teachers' professional competencies in writing test papers and English exercises. They found eight professional competencies improved by collaborative test construction (CTC), including testing knowledge, linguistic competence, intercultural communicative knowledge, curriculum resource utilization, reflection, computer information technology, pedagogy, and a two-way specification table.

Through semi-structured interviews, Farhady and Tavassoli (2021) examined EFL teachers' perceptions and application of language assessment knowledge (LAK) in a qualitative study. The study found no significant differences between teachers with high LAK and those with low LAK and no meaningful relationship between LAK level and students' performance on classroom achievement tests. However, teachers with high LAK wrote more extended tests with more varied sections and tasks.

Mohammadkhah et al. (2022) created a Language Assessment Literacy (LAL) model for Iranian EFL teachers, revealing four components: technical skills assessment, assessment purposes and principles, language pedagogy assessment, and scoring and interpretation of results. The findings suggest teachers' conceptions, knowledge, and beliefs influence their assessment practices and performance. The model can inform teacher education programs and professional development workshops, helping EFL teachers achieve LAL core competencies.

Thoyyibah (2022) examined the development of English teachers' reading tests in vocational schools using a case study approach. The study focuses on three English teachers in three schools and uses interviews and expert standards to understand the process. The results show that the first test is poor and good, with passing scores of 69 and 87, while the second test is inadequate, with scores of 51 and 56. The third teacher's reading test score is 81, indicating very good performance. Other teachers' abilities in test creation have not been effectively utilized in test design, relevance, balance, efficiency, validity, reliability, adequacy of test items, and technical voice of reading tests.

Zhao and Zhao (2023) conducted a study in China on involving learners in co-constructing assessment criteria in language classrooms. They

argued that using instructor-provided criteria could lead to discrepancies between assessment, teaching, and learning. The study involved two writing instructors and 146 tertiary students adapting assessment criteria in local contexts. The results showed that co-constructing criteria improved quality and developed learners' cognitive and metacognitive knowledge.

8. Methodology

The data was collected using a descriptive research design. Fraenkel and Wallen (2009) state that a descriptive design "summarizes the characteristics (abilities, preferences, behaviors) of individuals, groups, or (sometimes) physical environments (such as schools)" (p.14).

8.1. Population and research sample

This study comprises all Saudi EFL teachers in primary, intermediate, and secondary schools in Madinah Region during the first and second semesters of the 2021 and 2022 academic years. The study sample includes 227 EFL teachers: male (n=83) and female (n=144). The characteristics of the sample are described based on their gender, qualifications, and years of teaching experience as follows:

Table 1. Characteristics of the research sample

Variables		Frequencies N = 227	Percentages	Total (100%)
Gender	Male	83	36.6%	100%
	Female	144	63.4%	
Qualifications	BA in Education	67	29.5%	100%
	BA with an educational Diploma	108	47.6%	
	Master	46	20.3%	
	PhD	6	2.6%	
Years of experience	Less than 5 years	27	11.9%	100%
	6-10 years	72	23.8%	
	11-15 years	49	21.6%	
	16-20 years	54	31.7%	
	More than 20 years	25	11.9%	

As shown in Table 1, more than half of the participants, 144 (63.4%), were female teachers, and 83 (36.6%) were male teachers. About half of the

participants, 108 (47.6%), have a bachelor's degree and an educational Diploma. Also, 54 (31.7%) had 16 to 20 years of experience, whereas 72 (23.8%) had 6 to 10 years of experience, and 49 (21.6%) had 11 to 15 years of experience.

8.2. Instrumentation

The researchers designed a closed-ended questionnaire to the perspectives of Saudi EFL teachers on the practical criteria for test construction based on the reviewed literature (Ramírez, 2020; Swaie, 2023; Thoyyibah, 2022; Trisanti, 2019). The questionnaire is divided into three parts. The first part consists of questions about the sample demographic information. The participants were asked to answer two closed questions in the second part.

The first question was about perceiving knowledge-based language test construction and ranked using a five-point Likert scale (5= Excellent, 4= Very good, 3= Good, 2= Poor knowledge, and 1= No knowledge). The second question was about test question types that serve EFL teachers' language learning goals and ranked using a five-point Likert scale (5= Always, 4= Usually, 3= Sometimes, 2= Rarely, and 1= Never). Lastly, the third part consists of several items to measure teachers' perspectives on the practical criteria for test construction. The 32 items were ranked using a five-point Likert scale (5= Strongly Agree, 4= Agree 3= Neutral, 2= Disagree, and 1= Strongly Disagree). The items were also distributed on the seven central dimensions. They include clarity (7 items), relevancy (4 items), validity (4 items), reliability (5 items), flexibility (4 items), fairness (5 items), and feedback (3 items).

The Likert 5-point scale was employed to evaluate the variables, reflecting the extent of performance from low to high. The scales employed in the research instrument have five ranges: 1-1.80 for strongly disagree, no knowledge, and never, 1.81-2.60 for disagree, poor knowledge, and rarely, 2.61-3.40 for neutral, good, and sometimes, 3.41-4.20 for agree, very good, and usually, and 4.21-5 for strongly agree, excellent, and always.

8.3. Validity and reliability

Three language teaching experts and four experienced EFL teachers reviewed the questionnaire items and commented on their clarity, relevance, and adequacy to the purpose of the study and the dimensions. They suggested some linguistic and structural modifications and deleted two statements,

resulting in 32 items in the final version of the questionnaire. Moreover, a pilot study was conducted with 29 Saudi EFL teachers to compute the questionnaire's reliability and internal consistency to measure the extent to which the questionnaire items measure the same underlying construct.

Pearson correlation was employed to describe the strength and direction of the relationship between the questionnaire items and the items of each dimension. The correlation coefficient results indicated a positive, direct, and significant relationship between the total and 32 items with range values (r= .517**.287**) of clarity, (r= .615**.423**) of relevancy, (r= .551**.334**) of validity, (r= .475**.152*) of reliability, (r= .500**.349**) of flexibility, (r= .436**.229**) of fairness, and (r= .537**.414**) of feedback.

Cronbach's alpha coefficient was employed to measure all items' internal consistency in the questionnaire. The results indicated a high internal consistency and reliability between the 32 items (α=0.961). The reliability coefficient of the dimensions is as follows: clarity (α=0.969), relevancy (α=0.924), validity (α=0.961), reliability (α=0.921), flexibility (α=0.889), fairness (α=0.923), and feedback (α=0.984).

8.4. Data collection

On October 2nd, 2021, the questionnaire link was directed officially to the Department of Educational Supervision in Madinah, Saudi Arabia. It took about three weeks to receive 227 responses from EFL teachers.

8.5. Data analysis

The gathered data from the questionnaire was examined using descriptive statistics (frequencies, percentages, means, and standard deviations). Due to the nature of the research design and collected data, the Mann-Whitney U test was used as the primary inferential statistical method to address the third research question and test the hypothesis. Statistical analysis was performed using SPSS software (version 26).

9.Results

The research questions were answered using frequencies, percentages, means, and standard deviations, with the results of the questionnaire responses incorporating the gender variable.

9.1 Results of EFL teachers' knowledge of constructing language tests

Descriptive statistics were employed to answer the first research question, "How do Saudi EFL teachers perceive their knowledge about constructing language tests?"

Table 2. EFL teachers' knowledge about constructing language tests

Responses	Male (n=83)				Level	Female (n=144)				Level
	F	%	M	SD		F	%	M	SD	
Excellent	7	8.4	3.06	1.04	Good	29	20.1	2.91	1.34	Good
Very good	19	22.9				14	9.7			
Good	36	43.4				37	25.7			
Poor knowledge	14	16.9				43	29.9			
No knowledge	7	8.4				21	14.6			

Table 2 shows that 29 (80.6%) EFL female teachers perceive themselves as "excellent" in knowledge-based test construction, and 37 (50.7%) are "good" in knowing about test construction. In contrast, 43 (75.4%) EFL female teachers also perceive themselves as "poor in knowledge" about constructing tests, and 21(75%) indicated that they have "no knowledge" about designing tests. On the other hand, only 7 (25%) EFL male teachers perceive themselves as "lacking knowledge," and 19 (57.6%) of them are "very good" in having experience in constructing tests. Generally, males and females (M= 3.06 & 2.91) revealed "good knowledge" about test construction. The high response rate among Saudi EFL teachers can be attributed to their high qualifications, with 47.6% holding a bachelor's degree and 20.3% having a master's degree. Additionally, their years of experience in education, ranging from 16 to 20 years, are significant, with 23.8% having a 6 to 10-year background and 21.6% having 11 to 15 years. These qualifications and experience may enhance their test construction knowledge.

9.2 Results of EFL teachers' preferences of test question types

Descriptive statistics were employed to answer the second research question "answer the first research question, "How do Saudi EFL teachers perceive their knowledge about constructing language tests?"

Table 3. Test question types that serve Saudi EFL male teachers' language learning goals (n=83)

Test Types	F (%)							Levels
	Always	Usually	Sometimes	Rarely	Never	M	SD	
True-false	65 (78.3)	18 (21.7)	-	-	-	4.78	0.41	Always
Multiple-choice	65 (78.3)	18 (21.7)	-	-	-	4.78	0.41	Always
Matching	32 (38.6)	35 (42.2)	16 (19.3)	-	-	4.19	0.74	Always
Completion	32 (38.6)	35 (42.2)	16 (19.3)	-	-	4.19	0.74	Always
Error-correction	32 (38.6)	35 (42.2)	16 (19.3)	-	-	4.19	0.74	Always
Gap-filling	23 (27.7)	44 (53.0)	16 (19.3)	-	-	4.08	0.68	Always
Transformation	32 (38.6)	26 (31.3)	25 (30.1)	-	-	4.08	0.83	Always
Unscramble	20 (24.1)	-	53 (63.9)	6 (7.2)	4 (4.8)	3.07	0.71	Sometimes
Short essay	-	-	34 (41.0)	15 (18.1)	34 (41.0)	2.00	0.91	Rarely
Rewriting	-	-	17 (20.5)	35 (42.2)	31 (37.3)	1.83	0.75	Rarely
Open-ended	-	-	9 (10.8)	35 (42.2)	39 (47.0)	1.64	0.67	Never
Cloze items	-	-	8 (9.6)	33 (39.8)	42 (50.6)	1.59	0.66	Never
Translation	-	-	4 (4.8)	39 (47.0)	40 (48.2)	1.57	0.59	Never

The results in Table 3 show that EFL male teachers consistently use true-false, multiple-choice, matching, completion, error-correction, gap-filling, transformation, and transformation test types, with high mean scores ranging from 4.78 to 4.08. These results might be attributed to the nature of objective test items because they are considered reliable assessment methods

due to their ability to reduce workload, provide immediate feedback, and be faster and easier to grade. They are also suitable for specific tasks, allowing EFL teachers to test students on various language skills and statistically analyze individual performance.

In contrast, the subjective nature of test items like short essays, rewriting, open-ended, cloze items, and translation resulted in low mean scores ranging from 2.00 to 1.57. These results may be due to limitations, such as limited lesson material scope, difficulty in objective correction, low validity, low reliability, less representativeness, influenced by subjective elements, individualized assessment, and long correction time, making them unsuitable for EFL male teachers.

Table 4. Test question types that serve Saudi EFL female teachers' language learning goals (n=144)

Test Types	F (%)							Levels
	Always	Usually	Sometimes	Rarely	Never	M	SD	
True-false	96 (66.7)	41 (28.5)	7 (4.9)	-	-	4.62	0.58	Always
Multiple choice	95 (66.0)	41 (28.5)	8 (5.6)	-	-	4.60	0.59	Always
Error-correction	55 (38.2)	56 (38.9)	20 (13.9)	4 (2.8)	9 (6.3)	4.00	1.10	Always
Completion	36 (25.0)	33 (22.9)	35 (24.3)	21 (14.6)	19 (13.2)	3.32	1.35	Sometimes
Transformation	41 (28.5)	33 (22.9)	17 (11.8)	26 (18.1)	27 (18.8)	3.24	1.50	Sometimes
Gap-filling	32 (22.2)	38 (26.4)	28 (19.4)	16 (11.1)	30 (20.8)	3.18	1.44	Sometimes
Unscramble	-	32 (22.2)	95 (66.0)	10 (6.9)	7 (4.9)	3.06	0.70	Sometimes
Matching	38 (26.4)	27 (18.8)	8 (5.6)	26 (18.1)	45 (31.3)	2.91	1.64	Sometimes
Short essay	19 (13.2)	10 (6.9)	46 (31.9)	22 (15.3)	47 (32.6)	2.53	1.36	Rarely
Cloze items	23 (16.0)	14 (9.7)	3 (2.1)	45 (31.3)	59 (41.0)	2.28	1.48	Rarely

Test Types	F (%)					M	SD	Levels
	Always	Usually	Sometimes	Rarely	Never			
Rewriting	-	1 (.7)	36 (25.0)	57 (39.6)	50 (34.7)	1.92	0.79	Rarely
Open-ended	-	-	17 (11.8)	59 (41.0)	68 (47.2)	1.65	0.68	Never
Translation	-	-	5 (3.5)	67 (46.5)	72 (50.0)	1.53	0.57	Never

Similarly, Table 4 shows that EFL female teachers consistently use true-false, multiple-choice, and error-correction test types with high mean scores ranging from 4.62 to 4.00. On the other hand, they prefer completion, transformation, gap-filling, unscrambling, and matching test items with medium mean scores ranging from 3.32 to 2.9. Besides what has been mentioned above in Table...., the preference of EFL female teachers might be attributed to the reliability, speed, and ease of use of objective test items, which do not require language mastery, making them easier to implement in classroom-based tests.

9.3. Results of perspectives on the practical criteria for test construction

Descriptive statistics were computed to answer the third research question, "What are the perspectives of Saudi EFL teachers on the practical criteria for test construction?"

Table 5. EFL teachers' perspectives on the practical criteria for test construction

Dimensions	Criteria	Male (n=83)		Female (n=144)	
		M	SD	M	SD
Clarity	Write the objective of the test or exam.	4.57	.499	4.48	.591
	Provide pre-information on the conditions for the test.	4.43	.522	4.51	.542
	Choose test question items appropriate to content.	4.47	.570	4.52	.515
	Give clear, concise instructions for the whole test.	4.46	.548	4.57	.550

	Give comprehensible instructions for each question.	4.52	.526	4.49	.614
	Confirm that all the items are grammatically correct.	4.52	.503	4.49	.591
	Use appropriate language to students' proficiency levels.	4.54	.501	4.50	.555
Relevancy	Be sure that learning outcomes are reflected in all items	4.54	.501	4.62	.488
	Revise any redundancy in the test paper.	4.53	.549	4.50	.615
	Choose related questions to the content.	4.55	.500	4.62	.488
	Construct question items that permit students to demonstrate their knowledge of subject matter.	4.53	.502	4.58	.496
Validity	Construct question items measure learning objectives.	4.60	.492	4.62	.488
	Test what the students have learned only.	4.57	.522	4.40	.594
	Avoid combined questions in tests and exams.	4.52	.503	4.50	.591
	Engage students with different types of questions.	4.34	.547	4.50	.567
Reliability	Interpret the student's results after the tests and exams.	4.55	.500	4.71	.456
	Apply test techniques that have been worked with students.	4.53	.502	4.55	.499
	Design remedial materials for low achievers after tests.	4.51	.527	4.49	.567
	Confirm there is no room for teacher's personal interpretation.	4.57	.588	4.36	.696
	Respond to all students equally while monitoring them.	4.61	.514	4.51	.700
Flexibility	Make responses short by using key-specific words.	4.66	.524	4.56	.676

	Avoid clues that guide students to eliminate incorrect options or choose the correct ones.	4.65	.480	4.33	.811
	Offer extra marks to help students score high.	4.69	.467	4.58	.643
	Re-test students with language disabilities.	4.55	.649	4.50	.748
Fairness	Consider the individual differences.	4.66	.524	4.42	.725
	Include different types of questions.	4.59	.645	4.35	.864
	Make the student satisfied with item numbers.	4.63	.487	4.44	.782
	Use a systematic process for scoring each question test item.	4.64	.483	4.70	.459
	Ensure that there is only one correct answer per item.	4.60	.492	4.67	.555
Feedback	Allow students to express opinions on the given tests.	4.54	.501	4.61	.670
	Motivate students for their excellent grades.	4.60	.492	4.34	.969
	Give general comments on the tests and exams.	4.55	.500	4.59	.673

As shown in Table 5, the results reveal that both male and female EFL teachers have high mean scores in their responses regarding the practical criteria for test construction, indicating strong agreement on these criteria. These results might be attributed to the strong beliefs of EFL male and female teachers on test construction criteria, focusing on clear objectives, pre-conditions, appropriate question items, concise instructions, understandable questions, grammatically correct items, and proper language for students' proficiency levels.

In addition, The EFL teachers emphasize the importance of test relevance to learning outcomes, revising redundancy, selecting related questions to content, and creating question items that allow students to demonstrate their subject matter knowledge.

EFL teachers prioritize test validity and reliability, focusing on student learning, avoiding combined questions, engaging students with diverse questions, interpreting results, applying techniques, designing

remedial materials for low achievers, avoiding personal interpretation, and responding equally to all students while monitoring them.

Furthermore, EFL teachers prioritize flexibility and fairness in designing language tests, using key-specific words, avoiding clues, offering extra marks, re-testing students with disabilities, considering individual differences, including different types of questions, satisfying students with item numbers, using a systematic scoring process, and ensuring only one correct answer per item. Finally, EFL teachers prioritize providing feedback in designing language tests, allowing students to express opinions, motivating them for excellent grades, and providing general comments on tests and exams.

9.4. Results of the differences in the practical criteria for test construction

The Mann-Whitney U test was used to compare EFL male and female teachers' perspectives on test construction criteria and answer the fourth research question and test the research hypothesis, " There are no statistically significant differences at 0.05 between Saudi EFL male and female teachers' perspectives on the practical criteria for test construction."

Table 6. EFL teachers' perspectives on the criteria for test construction

Dimensions	Group	N	Mean	SD.	Mean Rank	Sum of Ranks	Mann-Whitney U	Z	Sig. (2-tailed)
Clarity	Male	83	31.51	2.72	113.47	9418.00	5932.000	-.094	.925
	Female	144	31.56	2.62	114.31	16460.00			
Relevancy	Male	83	18.16	1.74	111.14	9225.00	5739.000	-.516	.606
	Female	144	18.31	1.58	115.65	16653.00			
Validity	Male	83	18.02	1.65	113.06	9384.00	5898.000	-.168	.866
	Female	144	18.01	1.75	114.54	16494.00			
Reliability	Male	83	22.77	2.15	117.33	9738.00	5700.000	-.593	.553
	Female	144	22.63	1.96	112.08	16140.00			
Flexibility	Male	83	18.55	1.74	123.99	10291.50	5146.500	-1.826	.068
	Female	144	17.97	2.30	108.24	15586.50			
Fairness	Male	83	23.12	1.97	122.47	10165.00	5273.000	-1.532	.125
	Female	144	22.58	2.35	109.12	15713.00			
Feedback	Male	83	13.70	1.34	113.95	9458.00	5972.000	-.009	.993
	Female	144	13.54	1.80	114.03	16420.00			

Results of the Mann-Whitney U-test showed no significant differences between EFL male and female teachers in the seven dimensions: clarity, relevance, validity, reliability, flexibility, fairness, and feedback. The results of the p-values, which are larger than the 0.05 significance level, confirm that there are no gender differences in perspectives on the criteria of test construction.

10.Discussions

Analyzing the first two questions in the rating scale revealed that most EFL teachers perceive good knowledge about constructing language tests and prefer objective test question types that serve their language learning goals. A possible explanation for these results might be that EFL teachers' qualifications and teaching experiences influence their test construction knowledge, while subject matter and proficiency level in Saudi EFL learners may influence teachers' preferred test types. These findings agree with those obtained by Sun and Bin-Sihes (2020) who assured the EFL teachers' professional needs for test construction, including testing knowledge, linguistic competence, intercultural communicative knowledge, and curriculum resource utilization. Likewise, these results agree with Belarbi and Bensafa (2020), who suggested that EFL should design objective test items to tests with clear, concise instructions to avoid irrelevant questions. However, these results contradict Sun and Bin-Sihes (2020), who reported the complexity of writing test items, such as multiple-choice, true/false, matching, and reading comprehension, making EFL teachers construct less relevant tests to their learning outcomes.

Analyzing the responses for the seven dimensions, the results indicated strong agreement of EFL male and female teachers on their perspectives regarding the practical criteria for test construction. EFL teachers highly agreed on test relevance, validity, reliability, flexibility, and fairness in language test design. The findings may be attributed to EFL teachers being required to adhere to specific test construction criteria set by their supervisors and receive corrections during periodic visits. It might also be due to that experienced EFL teachers may have access to well-constructed test samples, which they share with colleagues formally or informally. These results are consistent with what Lemmetti (2015) reported regarding the importance of validity, reliability, and washback in constructing and using a test. Abahussain et al. (2020) also found that focusing on balanced, valid, and

reliable questions is an essential test construction criterion. Moreover, This result is consistent with Zhao and Zhao (2023), who ensured that using instructor-provided criteria improved test quality and developed learners' cognitive and metacognitive knowledge. However, this result contrasts Thoyyibah (2022), who revealed that EFL teachers' abilities in test creation have not been effectively utilized in test design, relevance, balance, efficiency, validity, reliability, adequacy of test items, and technical voice of reading tests.

The results also revealed that EFL teachers highly perceive the importance of reviewing redundancy, selecting related questions, and allowing students to demonstrate subject matter knowledge. They also consider avoiding combined questions, engaging diverse students, and providing feedback to motivate excellent grades.

These results support previous research on test construction criteria. For example, Abahussain et al. (2020) highlighted the importance of considering test criteria such as question variety, clear instructions, validity, and alignment with learning outcomes for the best test construction. In addition to the agreement of the model proposed Mohammadkhah et al. (2022) for practical test construction knowledge for EFL teachers. The most significant criteria for test construction are identified in four key components: technical skills assessment, assessment purposes and principles, language pedagogy assessment, and scoring and interpretation of results. However, these results contrast with those of Belarbi and Bensafa (2020), who found that the language test is not constructed to evaluate students' communicative or higher-order thinking skills, suggesting the need for practical EFL test criteria for language tests.

The results of the Mann-Whitney U-test revealed no significant gender differences in clarity, relevance, validity, reliability, flexibility, fairness, and feedback between EFL male and female teachers, confirming no gender differences in test construction criteria. A possible explanation for these findings is that both male and female Saudi EFL teachers agree on the significance of test construction in language teaching and learning. Another possible explanation for these findings is the cultural similarities that may influence their thoughts, values, and views on test design and development, potentially influencing their perception and actions. This result contrasts with

Farhady and Tavassoli (2018), who found that gender is a significant factor in EFL teachers' language assessment knowledge.

11.Conclusion & Recommendations

Test design is fundamental for assessing students' comprehension of course content and their application skills. Test construction, or test development, involves planning, designing, creating, administering, scoring, and statistically analyzing tests to ensure the validity of the resulting scores. Thus, the primary goal of the current quantitative study is to explore the perspectives of Saudi EFL teachers on test construction criteria and their effectiveness in teaching settings. The research also aims to understand their gender-specific views on test construction requirements. Therefore, the researchers developed a 32-5-point Likert scale questionnaire to measure teachers' perspectives. The sample includes 227 teachers, 83 males, and 144 females, during the 2021 and 2022 academic years. The results found that the test construction process is crucial for EFL teaching, learning, and assessment. EFL teachers generally have good test design and development knowledge, preferring objective items over subjective ones. They highly value the practical criteria of test construction and the essential role of all seven dimensions in the process. The Mann-Whitney U-test revealed no significant gender differences in clarity, relevance, validity, reliability, flexibility, fairness, and feedback perspectives among EFL male and female teachers.

The study contributes significantly to the literature on practical criteria for test construction, enhancing EFL teachers' understanding and making it valuable for school teaching and learning in EFL. It also contributes to the growing body of literature on test construction. Based on the findings of this study, the researcher suggests some practical recommendations. First, based on the recommendations of Alfallaj and Al-Ahdal (2017) and Ahmed's (2018), in-service English teachers should receive comprehensive training in practical test construction skills through a comprehensive program. They should be trained in designing integrated language skills-based exams and authentic assessments, which are recommended methods in the curriculum. Second, in-service teachers should be encouraged to attend conferences related to testing and catch up with recent trends in language assessment to gain assessment literacy. The Ministry of Education (MOE) should establish regular workshops and

seminars to build in-service teachers' capabilities in using new test construction methods. Third, the MOE should conduct regular workshops and seminars to enhance in-service teachers' test construction skills and collaborate with language teachers to create clear exam evaluation policies. Fourth, undergraduate teacher programs should focus on pre-service teachers, including courses on test strategies and skills.

More research is needed to replicate the current study with EFL teachers and students to generate more profound insights into their perspectives on the practical criteria for test construction. Researchers should explore the relationship between implementing well-designed classroom-based language tests and other variables such as student satisfaction and self-improvement. Moreover, it is highly recommended that the impact of test construction criteria on students' language achievement be investigated further.

12.Limitations of the study

This study has resulted in several significant findings. However, this study has only provided data on EFL teachers; a broader study that includes the views and concerns of both EFL students and teachers could be conducted to add to the body of research in this area. Furthermore, the study was conducted only among Saudi EFL teachers in some Saudi schools. Hence, further studies in Arab countries are encouraged to generalize the results on a broader scale.

Acknowledgments

The researcher would like to thank the Saudi EFL teachers who participated in this study for their prompt responses throughout the submission period. Thanks go to some Saudi educational supervisors for their administrative assistance during data collection.

References

Abahussain, M. O., Iqbal, M., & Khan, I. (2020). Standardization of EFL undergraduate skill exam papers: A case study at Majmaah University. *Arab World English Journal (AWEJ)*, *11*(4), 363–381. https://doi.org/https://dx.doi.org/10.24093/awej/vol11no4.24

Abbas, A. A. (1994). *Evaluating the assessment process in the EFL teaching programs and the general secondary education certificate English exams for 1989-1993*. University of Massachusetts Amherst.

Ahmed, K. A. (2018). Training in-service teachers in test construction skills (TCS). *King Khalid University Journal for Humanities*, *5*(2), 49–69.

Al-Seghayer, K. (2014). The actuality, inefficiency, and needs of EFL teacher-preparation programs in Saudi Arabia. *International Journal of Applied Linguistics and English Literature*, *3*(1), 143–151. https://doi.org/10.7575/aiac.ijalel.v.3n.1p.143

Alfallaj, F., & Al-Ahdal, A. (2017). Authentic assessment: Evaluating the Saudi EFL tertiary examination system. *Theory and Practice in Language Studies*, *7*(8), 597–607. https://doi.org/10.17507/tpls.0708.01

Algarabel, S., & Dasí, C. (2001). The definition of achievement and the construction of tests for its measurement: A review of the main trends. *Psicológica*, *22*(1), 43–66.

Alghamdi, F. (2016). Self-directed learning in preparatory-year university students: Comparing successful and less-successful English language learners. *English Language Teaching*, *9*(7), 59–69. https://doi.org/10.5539/elt.v9n7p59

Ali, S. R., Ahmad, H., & Khan, A. A. (2019). Testing in English language teaching and its significance in EFL contexts: A theoretical perspective. *Global Regional Review (GRR)*, *IV*(II), 254–262. https://doi.org/10.31703/grr.2019(IV-II).27

Alokozaya, W. (2022). Students' perception of alternative assessment: A qualitative meta-analysis. *International Journal of Curriculum and Instruction*, *14*(2), 1419–1441. https://orcid.org/0000-0000-0000-0000

Alqahtani, S. M. A. (2019). Teaching English in Saudi Arabia. In *In C. Moskovsky & M. Picard (ed), English as a foreign language in Saudi Arabia: New insights into teaching and learning English* (pp. 120–137). Routledge is an imprint of the Taylor & Francis Group. https://doi.org/10.1017/S0305741000010158

Alrzini, J., Pennington, D. R., & Dunlop, M. D. (2022). The impact of test elements on students' performance in EFL assessment. *In EFL. In 16th International Conference on Interfaces and Human Computer Interaction (IHCI 2022).*, 1–10.

Altheyab, A. H. (2023). Secondary school teachers' perceptions of the qualities of effective EFL teachers in Saudi Arabia. *American Journal of Education and Technology (AJET)*, *2*(3), 28–39.

Arslan, R. Ş., & Üçok-Atasoy, M. (2020). An investigation into EFL teachers' assessment of young learners of English: Does practice match the policy? *International Online Journal of Education and Teaching (IOJET)*, *7*(2), 468–484.

Bachman, L. F. (2000). Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing*, *17*(1), 1–42. https://doi.org/10.1191/026553200675041464

Belarbi, F. M., & Bensafa, A. (2020). An evaluation of the Algerian EFL baccalaureate exam under the cognitive domains of Bloom's Taxonomy. *Arab World English Journal (AWEJ)*, *11*(4), 534–546. https://doi.org/https://dx.doi.org/10.24093/awej/vol11no4.34

Caldwell, J. S. (2008). Comprehension assessment: A classroom guide. In *Insights into Second Language Reading*. The Guilford Press, New York.

Clay, B. (2001). *Is This a Trick Question? A short guide to writing effective test questions*.

Dikli, S. (2003). Assessment at a distance: Traditional vs . alternative assessments. *The Turkish Online Journal of Education Technology*, *2*(3), 13–19. http://www.tojet.net/articles/v2i3/232.pdf

Dollah, S., & Atmowardoyo, H. (2022). The obstacles faced by EFL teachers in e-assessment of students in online based learning. *PINISI: Journal of Art, Humanity & Social Studies*, *2*(5), 10–19.

El-Hassan, D. F. A. M., & Ahmed, R. B. A. (2023). EFL teachers' challenges on the usage of online assessment methods. *International Journal of English Language Teaching*, *11*(5), 1–11. https://doi.org/https://doi.org/10.37745/ijelt.13/vol11n5111

Evers, W. M., & Walberg, H. J. (2005). *Testing student learning, evaluating teaching effectiveness*. Hoover Inst Press Publication. https://doi.org/10.5860/choice.42-2940

Farangi, M. R., & Rashidi, N. (2022). The relationship between Iranian EFL teachers' conceptions of assessment and their self-efficacy. *International Journal of Language Testing*, *12*(2), 59–75. https://doi.org/10.22034/IJLT.2022.157125

Farhady, H., & Selcuk, M. (2022). Classroom-based diagnostic assessment practices of EFL instructors. *Iranian Journal of Language Teaching Research*, *10*(2), 77–94.

Farhady, H., & Tavassoli, K. (2018). Developing a language assessment knowledge test for EFL teachers: A data-driven approach. *Iranian Journal of Language Teaching Research*, *6*(3), 79–94.

Handrayani, D. (2022). Teachers' problems and challenges in conducting online assessment. *In A. Ben Attou, M. L. Ciddi, & M. Unal (Eds.), Proceedings of ICSES 2022-- International Conference on Studies in Education and Social Sciences (Pp.1-13), Antalya, Türkiye. ISTES Organization.*, *4*, 1–13.

Henning, G. (2012). Twenty common testing mistakes for EFL teachers to avoid. *English Teaching Forum*, *3*, 33–40.

Hidri, S. (2020). New challenges in language assessment. In S. H. (ed.) (Ed.), *Changing Language Assessment* (pp. 3–22).

Hirai, A., Oka, H., Kato, T., & Maeda, H. (2022). Development and validation of an English test measuring EFL learners' critical thinking skills. *Language Testing in Asia*, *12*(45), 1–22. https://doi.org/10.1186/s40468-022-00193-2

Imsa-ard, P. (2020). Voices from Thai EFL teachers: Perceptions and beliefs towards the English test in the national examination in Thailand. *Language Education and Acquisition Research Network Journal*, *13*(2), 269–287.

Işık, A. (2020). Do students feel that they are assessed properly ? *Iranian Journal of Language Teaching Research*, *8*(1), 63–92.

Ivanova, V., & Terzieva, T. (2016). Criteria for the construction of tests for language assessment and evaluation. *In Proceedings of the Doctoral Conference in Mathematics and Informatics MIDOC, Sofia, Bulgaria (Pp. 15-18)., October 2016*, 58–56.

Kopriva, R. J. (2008). *Improving testing for English language learners*. Routledge, Taylor & Francis.

Kustati, M., & Zurniati, V. (2019). EFL teachers' problems in constructing English mid-semester test at statejunior high school 12 padang. *Islamic Manuscript of Linguistics and Humanity*, *1*(1), 36–48.

Lemmetti, J. (2015). *What makes a good language test in EFL?* chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://gupea.ub.gu.se/bitstream/handle/2077/38440/?sequence=1

Mohammadkhah, E., Kiany, G. R., Tajeddin, Z., & Shayestefar, P. (2022). EFL teachers' assessment literacy: A contextualized measure of assessment theories and skills. *Language Teaching Research Quarterly*, *29*, 102–119. https://doi.org/10.32038/ltrq.2022.29.07

Ramírez, A. (2020). Challenges in the design and implementation of an English placement test for a Colombian Public University. *Gist Education and Learning Research Journal*, *21*(July-December), 191–208.

Saher, A.-S., Ali, A. M. J., Amani, D., & Najwan, F. (2022). Traditional versus authentic assessments in higher education. *Pegem Journal of Education and Instruction*, *12*(1), 283–291. https://doi.org/10.47750/pegegog.12.01.29

Sun, Q., & Bin-Sihes, A. J. (2020). An analysis on EFL teacher professional competences required for test construction: Based on the activity of a collaborative test construction in China. *Journal of Advanced Research in Dynamical & Control System*, *12*(5 Special Isuue), 617–630. https://doi.org/10.5373/JARDCS/V12SP5/20201797

Sun, Y. (2022). A review and discussion of in-service EFL teachers' language assessment literacy level in junior high school in China. *Studies in Applied Linguistics & TESOL*, *21*(2), 65–80.

Swaie, M. (2023). Assessment purposes and methods used by EFL teachers in secondary schools in Jordan. *Frontiers in Education*, *8*, 1–10. https://doi.org/10.3389/feduc.2023.1192754

Thoyyibah, L. (2022). Teachers' competence in a reading test construction. *Indonesian EFL Journal (IEFLJ)*, *8*(2), 205–214.

Trisanti, N. (2019). Assessment literacy analysis on designing classroom language test. *Advances in Social Science, Education and Humanities Research (ASSEHR), Volume 188, UNNES International Conference on English Language Teaching, Literature, and Translation (ELTLT 2018)*, *188*, 126–130.

Xu, Y., & Brown, G. (2017). University English teacher assessment literacy: A survey-test report from China. *Papers in Language Testing and Assessment*, *6*(1), 133–158. https://doi.org/10.58379/uzon5145

Zhao, H., & Zhao, B. (2023). Co-constructing the assessment criteria for EFL writing by instructors and students: A participative approach to constructively aligning the CEFR, curricula, teaching and learning. *Language Teaching Research*, *27*(3), 765–793. https://doi.org/10.1177/1362168820948458