P-ISSN 2708-6453	Pakistan Languages and Humanities Review	Jan-Mar 2024, Vol. 8, No. 1
O-ISSN 2708-6461	https://doi.org/10.47205/plhr.2024(8-I)40	[441-458]

Pakistan Languages and Humanities Review
www.plhr.org.pk

RESEARCH PAPER

Validity Argument for Professional Language Assessment: Evaluating 'PPSC Test for Recruitment to the Post of Lecturer English'

[1]Memona Mujahid* and [2] Dr. Nighat Shakur

1. PhD Scholar, Department of English, International Islamic University Islamabad, Pakistan
2. Assistant Professor, Department of English, International Islamic University Islamabad, Pakistan

***Corresponding Author** Memona.phdeng206@student.iiu.edu.pk

ABSTRACT

The current research aims to evaluate the validity of the Recruitment Test for Lecturer English (TRLE) conducted by the Punjab Public Service Commission (PPSC). The test scores are used to allocate subject experts to teach English language at degree colleges in Punjab. The researcher follows a mixed-method approach focusing on quantitative data analysis of the five PPSC-TRLEs held between 2013 and 2022. The theories of language test validity informed the validity argument and guided the researcher to devise a tailored analytical framework. The quantitative data analysis and comparison of five tests reveal that it under-represents certain areas of language and exhibits inter-test and intra-test inconsistency. The inclusion of inappropriate and insufficient content leads to invalid test development and unfair decisions. The findings suggest that the test is not flawed for assessing candidates' language proficiency and predicting their language ability. The study recommends further research to investigate the reliability of the same test and recruitment tests for other subjects.

KEYWORDS Language Test, Lecturer English, PPSC, Professional Assessment, Recruitment, Validity Argument

Introduction

Language assessment basically measures the ability of a learner to use the language effectively and purposefully. The main objective of language assessment is to provide information about language user's proficiency and competence. Alan Davies (1990) defines language testing as "a measured concentration on language use and knowledge" (p. 9). Language Testing International (LTI: the US-based test-development organization) defines language testing as "a broad category of testing that assesses aspects of a person's ability to understand or communicate in a particular language." The pioneer researchers and theorists in language assessment (Henning, 1987; Davies, 1990; Brown, 2004; Hughes & Hughes, 2020) identify five major kinds of language test including proficiency, achievement, diagnostic, placement and aptitude. These language tests are used for a variety of purpose such as assessment of language in academic and professional contexts. In addition to assessing students' achievement, aptitude and performance, language tests are also used to select and recruit language teachers. Regardless of the context, language tests can be used to measure a person's language ability and skills such as required for a particular job role. American Council on the Teaching of Foreign Languages notes that language assessment uses a variety of instruments or techniques to gather information about the "ability to understand, speak, read and write" (ACTFL, 2012). Likewise, Cambridge Assessment English defines language testing as "the practice and study of evaluating the proficiency of an individual in using a particular language effectively. This assessment can include various language skills such as reading, writing, listening, and speaking." The definition from Educational Testing Service explains the purpose of language testing, "It encompasses the evaluation of linguistic abilities across different language skills and can be used for various purposes such as educational placement, certification, or immigration" (ETS, 2019). However, the ability in a language requires the skills for actual use of language not just learning the fabric of language (Leung, 2022). Language proficiency requires linguistic competence that includes theoretical and grammatical knowledge as well as communicative competence that refers to the ability of using language in actual situations. Assessment of the full ability of a language

While designing and using a language test, validity is the basic question. Validity scrutinizes a test for its purpose, content, relevance and construct. Grant Henning (1987) notes that the first and foremost consideration in selecting or developing a test is, "what is the test going to be used for" (p. 10). High stake tests require higher levels of validity because their purpose is also critical such as placement, recruitment, achievement and proficiency. The basic argument of test validation is that the whole test or any of its component parts should adequately and appropriately measure the language skill it is supposed to measure (Henning, 1987; Brown, 2004; Hughes & Hughes, 2020). Henning asserts on the primary question that is the consistency between the test content and goal. So, the agreement between test objective and content is the starting point for evaluating the validity of a language test. Gronulund (1998) defines validity as the extent to which the test provides useful and accurate scores for measuring the language ability. For example, a test conducted to measure the reading ability should not measure previous knowledge of language. Douglas Brown (2004) defines validity as the most important principle and criterion for the effectiveness of a test. If the test is not valid, it will not provide reliable results. Hence, the decision made on the basis of test scores are also invalid and unreliable. Validity of a test can also be taken synonymous to the term 'accuracy' defined by Hughes and Hughes (2020) as the "accurate measures of the test-takers ability" (p. 1). A language test is valid if it measures the intended ability by an accurate method. The current study has selected the recruitment test of PPSC for English lecturer. The main purpose of the test is to assess the test-takers' ability to teach English language courses at inter and degree level. In order to teach the language effectively, the teachers are required to possess English language proficiency, skills and knowledge. The most authentic definition of validity comes from American Psychological Association (APA), "Validity information indicates to the test user the degree to which the test is capable of achieving certain aims" (APA, 1954, p.13).

American Education Research Association (AERA) writes, "Validity information indicates the degree to which the test is capable of accomplishing certain aims" (AERA, 1955, p. 15). Similarly, the agreed upon definition of validity is "Questions of validity are questions of what may properly be inferred from a test score; validity refers to the appropriateness of inferences from test scores and other forms of assessment" (APA, AERA, & NCME, 1974, p. 25). Validity is the inherent trait of a test and refers to the use of test scores for logical and true interpretations as well as decisions (Giraldo, 2020). Fulcher & Davidson (2007) argue that the notion of validity suggests that the designed test is intended to measure 'something' that is 'real' and there must be a consistency between test intentions and actual use. Bachman (1990) defines content validity as content relevance and content coverage. The test must include the contents that are relevant to its purpose and use. Similarly, the contents must cover all aspects of the domain of knowledge or skills to be tested.

Language testing is an important aspect of applied linguistics but it is different from assessment in other educational subjects. Because language cannot be solely treated as a subject matter rather it is more like art and culture that are acquired. First language (L1) is acquired naturally and the second language (L2) is learned as a communicative skill and behavior. Davies (1990) differentiates language from other subjects in education on the basis of the native speaker, "we believe there is a native speaker of a language but not a 'native speaker' of chemistry" (p.10). The author argues that language testing has double requirements that it is about language and it is a test. Although some aspects of language are just like other educational subjects yet it has special requirements because language is a; skill, ability, behavior, part of biological, psychological and cognitive constructs. Hence, language assessment provides triple message about skills, knowledge and development (Davies, 1990). Language testing adopts different names and forms based on the purpose.

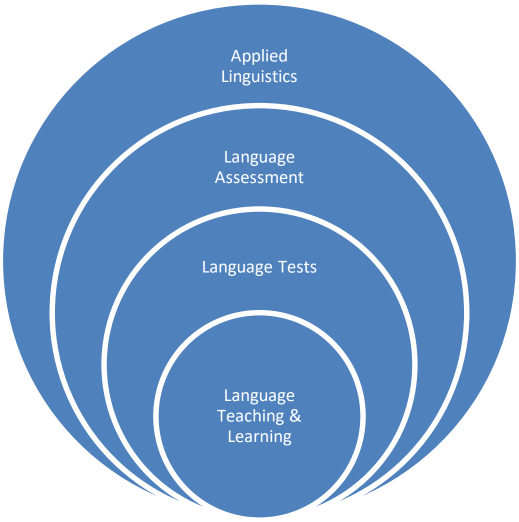

Figure 1. The relationship between applied linguistics, language teaching & learning (adapted from Alan Davies, 1990).

Language assessment and language development go side by side because the evidence from both these domains inform and improve each other. The data obtained from language assessment is widely used for research purposes such as to revise and devise the methods of assessment. The requirements for assessment and methods for testing evolve continuously to meet the demands of the language learners and users. Therefore, the validity of the test also needs constant judgment in order to keep the test aligned with its purpose and use. The issue of validity is crucial to consider in all testing situations, hence, the researcher has chosen the professional language test that is widely used in Pakistan to recruit professional English language teachers.

Validity standards were first proposed in 1954 and suggested four types of validity measurement including content validity, construct validity and criterion-related validity (Shepard, 1993). The measurement of validity depends on the use and purpose of the test. However, as a whole, all types of validity provide sufficient evidence on the authenticity of a test. Messick (1989) is regarded as a pioneer who gave the validity model largely based on the construct validity. He defines validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, p. 13). A language test not only provides information about the language ability but also leads to important decisions on the basis of the obtained scores, a flaw in validity can bring harmful consequences. One of the threats to test validity is inappropriate selection of content; the items do not match the objectives of the test (Henning, 1987). The content validity can be established by checking the items for their comprehensiveness and representativeness of their particular domain.

Basically, content validity is logical and intuitive lacking an empirical basis or coefficient. Brown (2004) reports, "There is no final, absolute measure of validity, but several different kinds may be invoked in support" (p. 22). The method to measure the validity of a test depends on its type, use and construct. Brown (2004) gives two criterion to check the basic validity of a test; (1) it actually tests the subject matter about which the conclusion are to be drawn (2) the test-takers actually perform the behavior that the test is seeking to measure. Similarly, Hughes & Hughes (2020) identify two main sources of test inaccuracy including test content and test techniques. Hughes and Hughes argue that a test need to be designed according to the skill it is going to measure. Moreover, the test is invalid if it contains insufficient items, e.g. the objectives or domain of language to be assessed are ten but the items are taken only from two or three domains. A test is not valid if it is purported to assess the reading comprehension skills of the test-taker but engages them in multiple choice questions requiring grammatical judgments. The conflict between the purpose of assessment and scope of the selected items leads to the issue of validity. For instance, accurate assessment of the writing ability of a candidate cannot be carried out by Multiple Choice test. Correspondingly, other language skills such as reading comprehension, speaking and listening cannot be accurately measured through MCQs test design. In high stake and professional test, the MCQs test technique is used for the purpose of convenience and economy. Because scoring a large number of compositions requires time and effort. However, the accuracy, validity and purpose of the test is sacrificed by choosing the incompatible technique. Thus, the scores obtained from the language assessment are not reliable enough to make decisions about the test-takers' language abilities such as proficiency level.

The content of the test should contain the representative samples of the language skill that is being assessed. A valid grammar test does not only measure knowledge of grammar but also assesses the understanding of correct grammatical structures. Hughes & Hughes (2020) argue that the validity of a test can be judged with reference to the specification of the sills or structures, "A comparison of a test specification and test content is the basis for judgments as to content validity" (p. 30). Validity is a very important aspect of the language test and it is directly linked with the accuracy of the assessment. High content validity cannot be achieved if the language areas highlighted in the test specification are under-represented or not represented at all through the items included. One of the greatest threat to content validity is the inclusion of the items and areas that are easy to test and score rather than important to assess (Hughes & Hughes, 2020). A test designed for recruitment of language teachers should mirror the kind of the tasks the teachers have to perform in an academic setting where the target audience is the English language learners (Malone & Montee, 2014).

Another prominent scholar, Sireci (2007) reports two very important arguments about validity of a test, "Validity is not a property of a test. Rather, it refers to the use of a test for a particular purpose… Evaluating test validity is not a static, one-time event; it is a continuous process" (p. 477). The author notes that the concept of content validity is simple and easier to understand, it is the fundamental consideration for a language test. Although validity has further categories yet it can be taken in a general sense as argued by Sireci, the author supports his claim by drawing inferences from Aera et. al (1999), "Validation can be viewed as developing a scientifically sound validity argument to support the intended interpretation of test scores and their relevance to the pro- posed use" (in Sireci, 2007, p. 478). The evidence for validity comes from five sources and one of them is the test content. However, validation of a test is not possible without referring to its purpose for which the test scores will be used. Furthermore, the author argues that the most relevant evidence for the validity of a test comes from testing purpose. An integration of theory and evidence is the basic step to measure the key validity of a test. Sireci (2007) suggests designing and prioritizing some effective probes to gather evidence for validity argument that can criticize the use of test scores. The researcher strongly favors an argument-based approach for validation and emphasizes the need to develop validity questions or criteria based on the fundamental purpose and use of test scores. The validity of a test cannot be established until the use of test scores is defined, in other words, the purpose of the conducting the test.

A valid test cannot be designed without proper pre-planning, validity is an integral part of test-development process and plays its role at each stage. Tyler (1934) gives a comprehensive view of test development and evaluation, he proposes that an educational test should be developed after establishing the objectives clearly. The specific focus should be on the actual skills and knowledge that test-takers' are expected to demonstrate. The items included in the test constitute a representative sample of the behavior or subject domain, "A fundamental assumption in all testing is that a sampling of student reactions will give a measure of his reactions in a much larger number of situations" (Tyler, 1989, p. 23). Therefore, the test must contain sufficient, accurate and appropriate items from each content domain that is being assessed. Otherwise, the validity and authenticity of the test and its scores is at stake. Based on research evidence and theories of validity, Shepard (1993) identifies five major steps in developing a valid test for professional assessment. The first step is to establish the appropriateness of content universe that is to determine the suitability of the test content for its purpose. This involves the basic questions like 'what knowledge or skills are essential'.

The test developers need to establish a balance between theoretical knowledge and practical application to assess the performance ability of the test-taker in a particular field. The researcher gives a particular example to understand his concept, "When judging a lawyer's competence, what is the proper balance between book learning and ability to mount an oral argument" (Shepard, 1993, p. 414). The second step is to assess the adequacy of content sampling obtained from the established content universe. This step is largely concerned with construct validity as it requires logical analysis instead of random or mechanical process to select the representative items. The tasks on the test must align with the domain specifications and the testing technique or mode of assessments should also be selected carefully. The third step demands expert judgment for the evaluation of content. The opinions from professionals and job experts together with evidence from job analysis and behaviors required for a particular job role are needed in this stage. The fourth stage is concerned with conceptual analysis of the internal elements, subdomains and the interrelationship of tasks. The underlying process affecting test performance and a clear rationale for test use are assessed in this stage. The fifth and last stage is integrating conceptual and substantive analyses with empirical studies for establishment of overall validity of the test. All major theories of validity suggest that the procedure of test development can also be applied for evaluating the validity of a test because an authentic test integrates the validity in the development process. Therefore, the stages of professional test development summarized by Shepard (1993) can be used as theoretical basis for test validation. After a comprehensive critical analysis of theories of validity, Shepard argues that the fundamental method to evaluate the validity of a test is to scrutinize the test under the light of some important questions, "What does the testing practice claim to do?... What are the arguments for and against the intended aims of the test? What does the test do in the system other than what it claims, for good or bad?" (p. 429). Scrutinizing an established test by asking these fundamental questions can yield significant information about the authenticity of the test. Passing a test through rigorous validity system is very essential, in fact, it is a continuous process.

Conducting the validity study of high-stake language tests is an established researched area. In this regard, Fulcher (1997) evaluates a placement test conducted by university of Surrey. The test is significant because it assesses the language proficiency of the students and places them in the respective category so that they can receive adequate support in language learning. The researchers investigates the administrative and logistic constraints as the usefulness of the test. The study concludes that the test meets its purpose despite minor issues in reliability and validity. Wolf et al. (2008) examines the validity of large scale tests developed by the US states in response to 'No Child Left Behind Act of 2001'. The researchers argues that the states have devised language assessment tests in a short time but they face issues in measuring the validity. The researcher provides a comprehensive overview of the validity issues, framework and the existing status of validation. This study also highlights the significance of validity for large scale tests.

Similarly, Winke (2011) conducted the validity research for a high-stake test the 'English Language Proficiency Assessment' (ELPA) largely used by the educational system of Michigan, USA. The researcher evaluates the perceived effectiveness of the test by taking teachers' and test administrators' reviews. Through qualitative analysis, the researcher derives three important themes that are not mentioned in the test document. So, the study emphasizes the significance of teachers' involvement in testing and suggests improvements in the instrument. Huang and Flores (2018) critically evaluate 'English Language Proficiency Assessment for the 21st Century' (ELPA 21) that is a standard test used at a large scale in eight states. The purpose of this online test is to place students at appropriate academic and career level. The researchers investigate the test completely checking its reliability, validity, authenticity, washback effects, practicality, bias and fairness. The study concludes that the test is fully valid and accurate for its purpose but a lack of validity document causes some doubts on its quality. The study indicates the significance of providing validity evidence for high-stake tests that are used for important decisions. International English Language Test System (IELTS) is a widely acknowledged instrument for language assessment across the world. Hashemi and Daneshfar (2018) carry out their study to evaluate the validity, reliability and washback of IELTS. The researchers report some issues in test reliability and suggest revision. Furthermore, they propose focusing on washback to improve test usefulness and authenticity. A brief overview of these studies emphasizes the importance of establishing validity measures for language tests used in educational as well as professional contexts.

Nationally designed language tests are widely used in Pakistan for various purposes such as students' placement in educational programs as well as for language assessment of professionals. The most reputable and high-stake tests include GAT (Graduate Assessment Test-subject and general), HAT (HEC Apptitude Test for scholarship), NTS, PPSC and FPSC recruitment tests. Almost all of these tests contain a section to measure test-takers' language ability. However, these tests are being used for many years and give no information on their validity or reliability. The review of the previous literature has yielded very few results, only one study is worth-mentioning that was conducted in Pakistan regarding the feasibility of tests in primary school education system. Andrabi et al. (2002) conducted a comprehensive feasibility survey to evaluate the quality of tests in schools. The researchers analyzed the content, rationale and administrative practices to judge the test validity. The study concludes that language tests need wider range of items to be valid like mathematics that indulges students in extensive practice and performance. The lack of validity evidence for national tests demands extensive research, therefore, the current study has selected one of the most important and high-stake language test for the validity argument. The next section of the paper elaborates the research process in detail.

Material and Methods

The present study applies a mixed method approach for analyzing the data obtained from the five selected samples. However, it largely relies on quantitative approach for data analysis. The data is represented through tables and pie charts for a clear representation and comparison. Furthermore, the quantitative findings are elaborated and interpreted through qualitative approach. The probes designed for theoretical framework have been answered through quantification. While the main research questions have been answered through qualitative interpretation.

Theoretical Framework

Kane (1992) proposed an argument-based approach for test validation that offers a simple and systematic process for creating a link between evidence and use of a language test. For the purpose of current study, the evidence comes from the test data-the PPSC recruitment test for English lecturer. While the use of the test is assessment of test-takers' language proficiency and competence which is required for teaching English as a second language in Pakistani context. Kane builds his claim on the basis of interpretative argument and validity argument. The interpretative argument specifies claims or inferences about the intended meaning and use of test scores- recruitment for the selected test. While the validity argument seeks support from empirical evidence. Another language testing expert, Lewkowicz (2000) asks a number of critical questions about the authenticity of language test designed for targeted language use. The particular concern of this paper is the probe, "To what extent can/do test tasks give rise to authentic-sounding output which allow for generalizations to be made about test takers' performance in the real world?" (p. 51). The researcher has followed the general theories of validity but specially focused on Kane's argument and Lewkowicz's probes that led her to the development of a tailored framework for building a validity argument for PPSC-TRLE. The purpose of the PPSC recruitment test is to assess the test-takers' ability for teaching English language courses at degree colleges of Punjab. The test scores are used to determine the merit criteria and to recruit the test-takers as lecturer in English. Hence, the specific purpose of the test is to assess candidate's language proficiency and knowledge that is required to teach ESL (English as a Second Language) learners enrolled in intermediate and degree programs.

Research Design

Based on the theories of validity, the researcher has developed a validity argument for the evaluation of the test. Moreover, the researcher has developed a specialized framework and criteria to test the fundamental validity of the PPSC test for recruitment lecturer English. The framework has analyzed the domains of language covered in the test and put them in relevant categories. This step identifies the specifications of the test and areas of language it assesses. In the next step, the number of items from each area is identified. Then, the test specifications are validated for its purpose. The test is validated on the criteria provided by the fundamental questions on scope, depth, balance, consistency and quantity of items in the test. The analysis provides responses for these questions and the inferences drawn from the results are used to answer the main research questions.

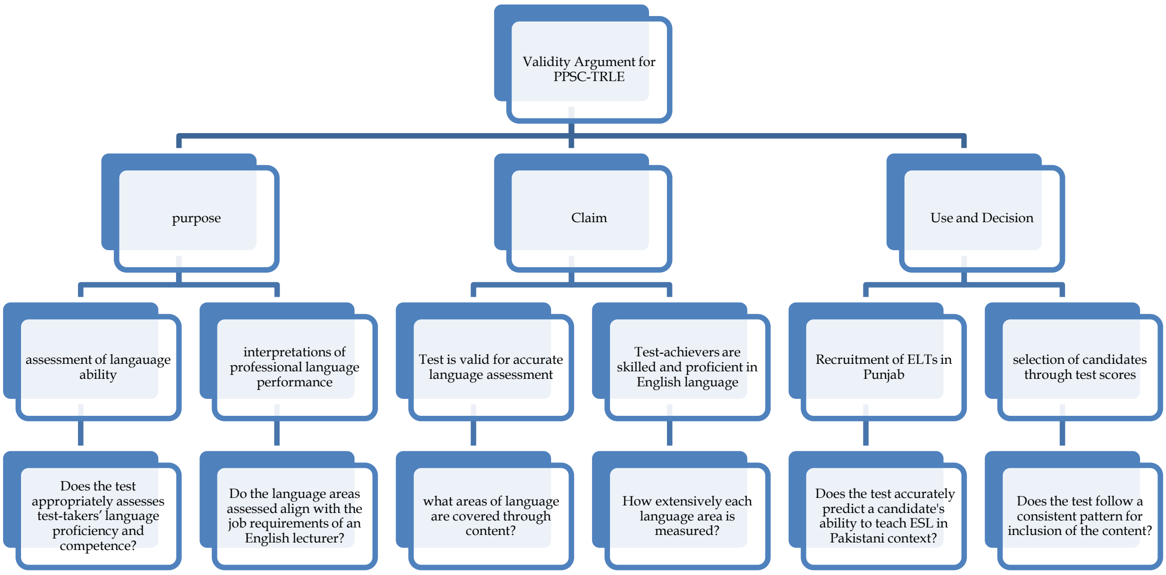
Validity Argument for PPSC-TRLE
purpose
Claim
Use and Decision
assessment of langauage ability
interpretations of professional language performance
Test is valid for accurate language assessment
Test-achievers are skilled and proficient in English language
Recruitment of ELTs in Punjab
selection of candidates through test scores
Does the test appropriately assesses test-takers' language proficiency and competence?
Do the language areas assessed align with the job requirements of an English lecturer?
what areas of language are covered through content?
How extensively each language area is measured?
Does the test accurately predict a candidate's ability to teach ESL in Pakistani context?
Does the test follow a consistent pattern for inclusion of the content?

Figure 2. PPSC-TRLE: The Research Design and Procedure for Validity Argument

Probes for Validity Argument

Theories of validity particularly Kane's theory of validity argument and Lewkowicz critical inquiry provide basis for the development of certain probes, the researchers has investigated the general validity of the selected test in the light of the five critical questions. The responses obtained from these probes have been interpreted to answer the main research questions.

1. Which areas of language does the test cover?
2. How extensively the included items cover each language area?
3. Does the test follow a consistent pattern for inclusion of the content?
4. Does the test appropriately assesses test-takers' language proficiency and competence?
5. Do the language areas assessed align with the job requirements of an English lecturer?

Data Collection

The data for the present research includes the written test for the post of lecturer English held by Punjab Public Service Commission (PPSC) almost every two to three years. The mentioned test is the main assessment criteria for the recruitment of the candidates who have completed minimum 16 years of education in the subject of English.

Sampling

For the current study, the researcher has selected 5 tests held in 2013, 2015, 2017, 2020, and 2022. The tests have been selected through purposive sampling technique. These tests were used over the past 10 years to recruit English lecturers at degree colleges throughout Punjab. The rationale for choosing 5 tests is to validate the validity argument on the basis of sufficient samples. Moreover, a comparative analysis of the test contents has informed about the general test pattern and trend of content inclusion.

Data Source

The data has multiple sources that include books, internet websites and social media. All the past papers/test of PPSC for the recruitment of lecturer English are available in test-preparation books sold in the local market. Another source to access the papers are various websites on the internet that have uploaded the data publicly. The third source is social media, especially Whatsapp and Facebook groups as well as pages. The researcher has retrieved the original tests from the internet sources. The data is widely circulated throughout country and is available publicly.

Delimitation

The test contains a total of 100 items each carrying one number/score. Moreover, the test has three major sections including literature, linguistics and general knowledge or current affairs. The items representing literature and linguistics are distributed randomly throughout the test, it does not identify each section separately. However, for the purpose of analysis, the researcher has picked up only the items that assess test-takers' language ability. Analyzing the questions asked about English literature is not the scope of the test because the research aims at arguing the validity of the test in the area of linguistics.

Analysis and Interpretation

The first question has been answered after the complete scrutiny of each test. The researcher has identified the items that represent the area of language and linguistics. Then, the researcher has placed each item in the relevant category. The number of items per category has been identified to quantify the representative sample.

Probe 1. Which areas of language does the test cover?

The test covers four main areas of language and linguistics including:

1. Pure and Applied Linguistics
2. Theoretical Knowledge of English Language
3. English Language Skills and Proficiency
4. English Language Teaching (ELT)

These four areas are further specified for the purpose of analysis, the researcher has identified the particular sub-area that each item on the test represents.

Pure and Applied Linguistics

The test includes representative items from the following 10 sub-areas of pure and applied linguistics:

i. Phonetics and phonology
ii. Morphology
iii. Semantics
iv. Syntax

v. Pragmatics
vi. Semiotics
vii. Language and history of language
viii. Theories of language acquisition and learning
ix. Theories of Linguistics
x. Sociolinguistics

Theoretical Knowledge of English Language

This domain is further divided into two categories and each one has sub-areas:

Lexical Knowledge

i. Vocabulary

ii. Idiomatic Expressions

Grammatical Knowledge

i. Sentence types

ii. Tenses

iii. Nouns/pronouns

iv. preposition

v. adverb

vi. adjective

English Language Skills and Proficiency

i. Reading comprehension through sentence completion

English Language Teaching (ELT)

i. Methods in English Language teaching

Probe 2. Does the test follow a consistent inter-test and intra-test pattern for assessing language areas through content selection?

Inter-test refers to the consistency between the five different test papers while intra-test refers to the consistency of language areas within the same test. The quantitative comparison of the 5 selected samples suggests that the PPSC-TRLE does not follow a consistent pattern for inclusion of specific areas of language. Moreover, the test does not represent clearly defined and fixed language areas for which the test-takers are assessed. The analysis reveals that on each test, areas of language to be assessed are selected randomly. The following pie charts represent the percentage of 4 language areas on each test.

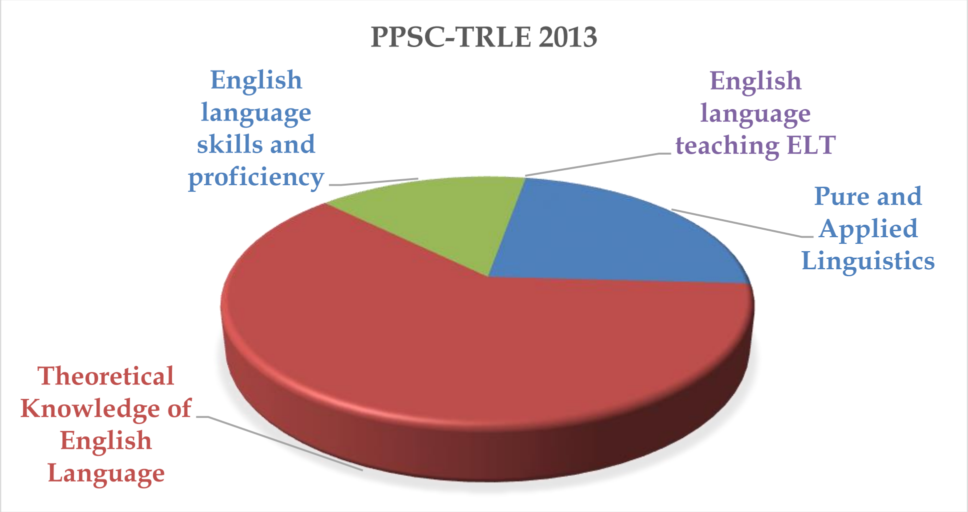

Figure 3. Comparative ratio of representative items from 4 language areas-2013

The PPSC-TRLE 2013 assesses theoretical knowledge of English language more widely than other 3 areas, zero number of items have been found from the area of ELT. The content selection is inconsistent and the comparative ratio of each area differs greatly.

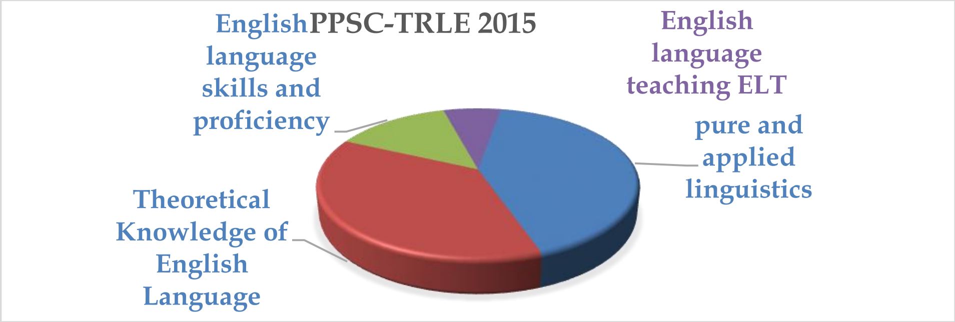

Figure 4. Comparative ratio of representative items from 4 language areas-2015

Figure 4.2 illustrates that the PPSC-TRLE 2015 has almost equal ratio of items from two language area. However, linguistics and grammar excels over other two areas. But, the consistency is again compromised.

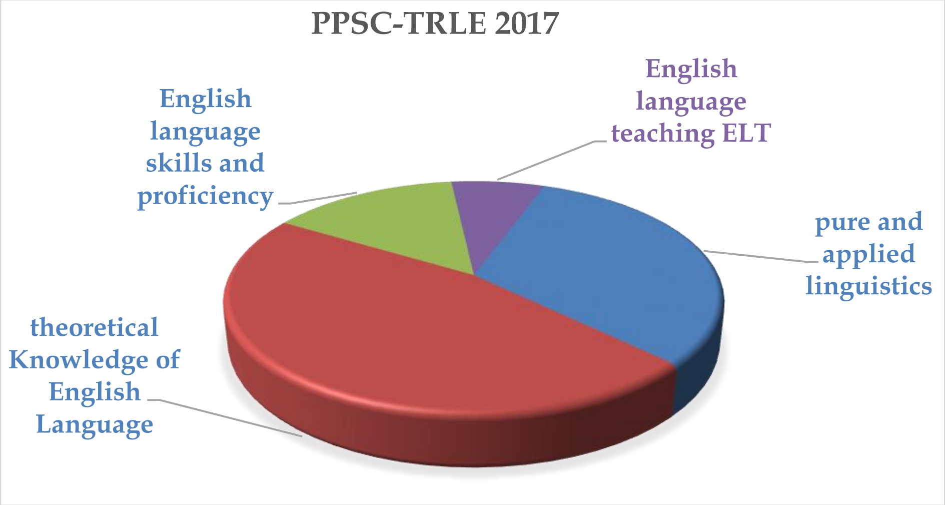

Figure 5. Comparative ratio of representative items from 4 language areas-2017

One language area dominates over other three representing intra-test inconsistency. The assessment of language proficiency has not been addressed fully. Likewise, ELT is also an under-represented area.

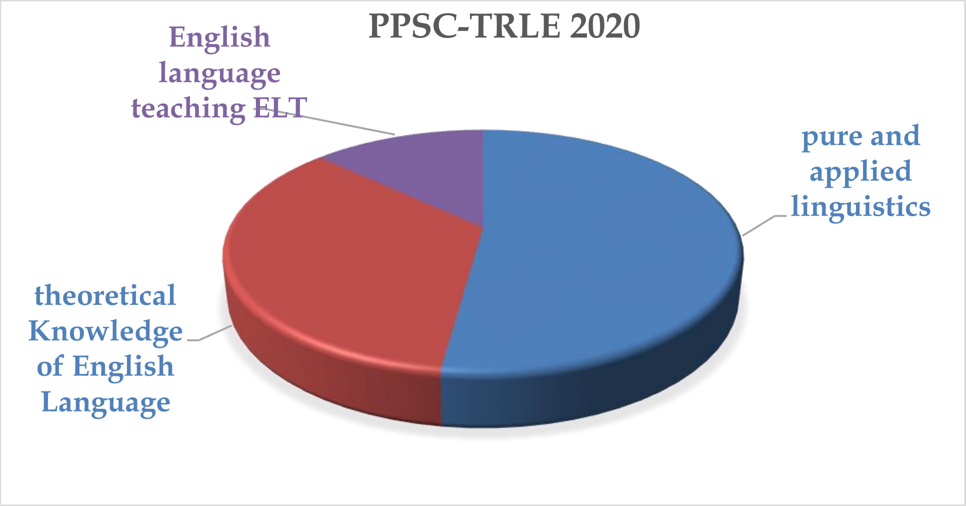

Figure 6. Comparative ratio of representative items from 4 language areas-2020

Test number 4 does not include any item to assess candidates' English language proficiency and skills. Figure 4 shows that this sample has high level of inter-test and some intra-test inconsistency as it evades the previous trend by giving no significance to one major areas of language assessment.

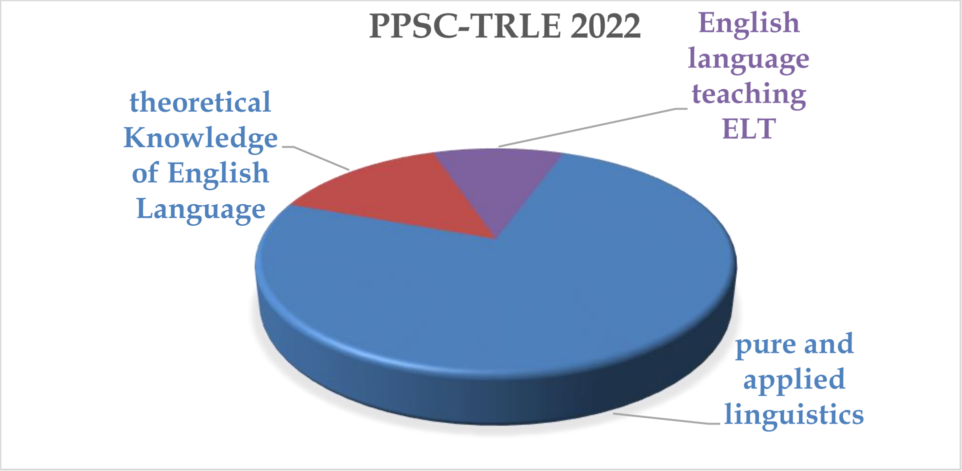

Figure 7. Comparative ratio of representative items from 4 language areas-2022

The figure 4.5 demonstrates that this test sample has also surpassed the previous trend by missing the items from the area of English language proficiency. Moreover, it largely focuses on pure and applied linguistics thus under-representing other two areas.

Probe 3. How extensively the included content items represent and assess each language area?

The answer of this question for validity argument lies in calculating the number of items representing each sub-area of language. The researcher has picked up items relevant to language and linguistics and then placed them under the respective categories. The following tables give a quantitative comparison of the five selected test samples.

Category 1 Pure and Applied Linguistics

Table 1
Number of representative items from 10 sub-areas of Pure and Applied Linguistics on each test

Sr.#	Sub-areas	2013	2015	2017	2020	2022	Average
1.	Phonetics and phonology	1	4	3	4	3	3
2.	morphology	1	2	1	0	3	1.4
3.	semantics	0	0	1	2	2	1
4.	syntax	1	4	1	1	1	1.6
5.	pragmatics	0	1	0	1	1	0.6
6.	semiotics	0	0	1	0	1	0.4
7.	Language and History of language	2	0	1	0	2	1
8.	Theories of language acquisition and	0	0	3	3	2	1.6
9.	Theories of Linguistics	0	3	2	0	2	1.4
10.	Sociolinguistics	1	4	0	1	4	2
	Sum total	6	18	13	12	21	14

The distribution of items from each area is inconsistent across 5 tests, some language areas are under-represented and some are not represented in the selected samples. Inclusion of only one items from each area is not sufficient for assessing test-takers' knowledge in the respective area.

Category 2. Theoretical Knowledge of English Language

Table 2
Number of representative items from 8 sub-areas of theoretical knowledge of English language

Sr.#	Sub-areas	2013	2015	2017	2020	2022	Average
A.	**Lexical Knowledge**						**7.2**
1.	vocabulary	10	10	10	3	1	6.8
2.	Idiomatic expressions	0	0	0	1	1	0.4
B.	**Grammatical knowledge**						**1.4**
3.	Sentence types	0	0	2	1	0	0.6
4.	tenses	0	1	2	1	0	0.8
5.	Nouns/pronouns	0	1	0	0	0	0.2
6.	preposition	6	3	4	2	0	3
7.	adjective	0	0	0	0	1	0.2
8.	adverb	0	1	1	0	1	0.6
	Sum total	16	16	19	8	4	12.6

The test assesses the lexical knowledge through vocabulary (synonyms and antonyms) and idiomatic expressions. The number of vocabulary items is consistent in three test-2013, 2015, 2017. However, the in the recent two tests (2020 and 2022), the vocabulary items are insufficient to measure test-takers' lexical knowledge. Likewise, the grammatical knowledge has been tested through sentence types, tenses and parts of speech. But, the number of items for each category are too low to give accurate information about test-takers' command over grammar. Moreover, the items are dispersed unequally across five tests referring to inconsistency in the test pattern and design. The test neglects some very important areas of grammar and parts of speech such as punctuation, proverbs, articles, conjunction, etc.

Category 3: English Language Skills and Proficiency

Table 3
Number of representative items from 1 sub-area of English language skills and proficiency

Reading comprehension and writing	2013	2015	2017	2020	2022	Average
sentence completion (MCQs)	4	6	6	0	0	3.2

According to the definitions of language proficiency, it refers to ability to perform through four skills that include reading, writing, listening and speaking. However, the PPSC-TRLE measures only writing and reading skill through sentence completion. The fragmented sentences missing a word or two cannot accurately measure candidate's language skills. The language assessment specialists (Henning, 1987; Brown, 2004; Hughes & Hughes, 2020) do not approve such type of content for assessing language proficiency. Therefore, the test establishes no significant validity for assessing candidate's proficiency that is a major requirement for English language teachers.

Category 4: English Language Teaching (ELT)

Table 4
Number of representative items from 1 sub-area of ELT

Pedagogy	2013	2015	2017	2020	2022	Average
Language teaching methodologies	0	3	3	0	0	1.2

The fourth area covered in the test is ELT that is another necessary skill required for effective language teaching. However, only 2 tests include 3 items from this area which assess candidate's general knowledge of teaching methodologies. This areas is also underrepresented and neglected in the test. The quantitative data represented in tables illustrate that insufficient number of representative items have been included from most of the sub-areas. The test content does not cover each language area sufficiently and extensively. Therefore, the quantitative argument shows that the test has a very low validity in terms of accurately measuring the language areas.

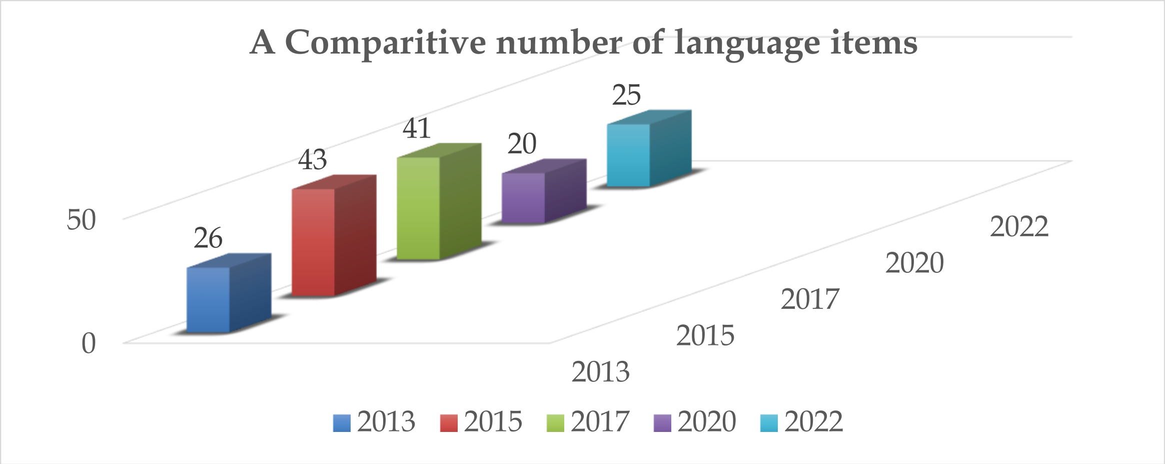

Figure 8. A comparison of the quantitative trend of language items (questions) on 5 tests

The comparative analysis reveals inconsistency in the number of total language questions on the test, the validity of the test is low in this regard. Hence, it shows both inter-test and intra-test inconsistency and needs to set standards for inclusion of specific number of language-assessing items on each test.

Probe 4. Does the test appropriately assesses test-takers' language skills?

Linguists agree that language skills include two major types of skills including receptive and productive. The receptive skills involve listening and speaking while productive skills refer to reading and writing. In order to be proficient and competent in a language, a user should possess all four language abilities or skills. Language skills and proficiency can be assessed through linguistic knowledge; the knowledge of grammar, language structure & system and linguistic competence; the ability to use language in real life situations or in a particular context. Researchers in language assessment argue that the test should contain the representative and valid sample of the content. For example, to test the reading and comprehension ability, the test should contain the tasks which accurately judge the reading skill of the test-takers. Likewise, writing skill requires the test-takers to produce a written composition or to choose correct lexical items to complete a passage.

Language Competence

In the light of these propositions by language experts, the selected test does not contain sufficient items to test the language skills. There are few items on the test which requires the test-takers to choose one of the four missing options in order to complete a sentence. Out of five selected samples, only three tests included some items that can assess reading comprehension skill, the average number of items per test is 3.2 that is insufficient to test the intended language area. Moreover, the scrutiny of the selected samples yield no results on other language skills; writing, listening and speaking. Therefore, the appropriateness of the test is not proven in assessing test-takers' language skills, proficiency and competence.

Language knowledge

The test includes items that assess test-takers' knowledge of language through grammar and vocabulary. Depending on the sub area they measure, the researcher has placed these items in two categories including lexical knowledge and grammatical knowledge. Almost each of the five selected samples contain representative items from this area of language. However, the number of representative items and the sub-areas of knowledge covered are inconsistent across the five samples.

Items assessing vocabulary through antonym, synonyms and pair of words exist in all samples but their number varies from 1-10. An average of 6.8 items per test can be considered as sufficient to assess test-takers' knowledge of English vocabulary. However, the number of items per test is inconsistent, only 3 test samples contain 10 items while the other 2 include one and three items, it raises questions about the accuracy of judgment. The test measures the knowledge of English language structure through grammatical rules and parts of speech. The average number of representative items of grammar is 1.4 and that of parts of speech is 4. The sum total of the average is 12.6 that represents the number of representative sample in five tests to assess test-takers' knowledge of English language.

Probe 5. Do the language areas assessed align with the job requirements for the lecturer in English language?

The EL teachers require sufficient competence and proficiency in English language to perform various job roles such as;

- Understanding and teaching the course content
- Imparting necessary language skills in students

- Teaching reading and writing skills
- Translating from English to Urdu and vice versa
- Teaching functional use of language and English for specific purpose (ESP)
- Communicating with students

However, the analysis of the five selected tests reveals that they do not cover all areas of language thus, ignore the assessment of language competency, proficiency and skills except reading comprehension that is measured through a small number of items. Moreover, each test under-represents the language areas it intends to measure. The number and average of items per language area are insufficient to measure the overall language ability of the potential English lecturers. The test largely assesses theoretical knowledge of language and linguistics and hardly focus on practical knowledge of English that demands actual use of language in classrooms.

Findings

A careful scrutiny of five selected PPSC-TRLE reveals that it cover four major areas of language including (1) pure and applied linguistics, (2) theoretical knowledge of English language, (3) English language skills and (4) English Language Teaching-ELT. The first area is touched almost from all aspects but the number of representative items are insufficient to give a full picture of test-takers' knowledge. Likewise, the theoretical knowledge of language structure and system has been assessed through tenses, sentence types, lexis, and a few parts of speech. The grammatical rules are not assessed fully but some tests include sufficient number of lexical items to judge vocabulary. The area of language skills is hardly assessed, only a few MCQs require the test-takers to pick up the right choice to complete a sentence. Hence, the test does not measure the four language skills. Similarly, another vital area is ELT that is the practical knowledge of teaching language in classrooms. However, the test also ignores this aspect by including only a few items across 5 selected samples. A comprehensive quantitative analysis of the test content, in the light of validity theories, reveals that it contains insufficient items and does not cover all areas of language. Moreover, the test also shows intra-test and inter-test inconsistency for assessing certain areas of language and excluding the others. Each area of language is not assessed extensively, it provides incomplete information about the test-takers' language ability. Hence, the test does not fully meets its purpose that is the adequate assessment of test-takers' language ability (knowledge, competence and performance) required for teaching English at degree colleges of Punjab. Moreover, the evidence from the analysis reveals that the test is not a fully authentic instrument to accurately predict a candidate's language ability to teach ESL in Pakistani context.

Conclusion

Validity is an integral part of a language test, it is an undeniable aspect that cannot be ignored at any cost. A compromise on the validity means getting unfair results from the scores that can adversely affect language teaching, learning and assessment systems. Therefore, no test should be used without passing through rigorous trial so that its purpose, content and use perfectly align with each other. The researcher suggests conducting further validity research to judge the validity of the test content whether it contains accurate and appropriate items for assessing each area of language. In this regard, opinions from a sufficient number of subject matter experts (SMEs) can be obtained to check the validity of each test item. The study concludes that the PPSC test for the recruitment of English lecturer requires careful scrutiny and thorough revision for making it a valid tool aligned with its purpose and an authentic predictor of candidates' English language ability.

References

Andrabi, T., Das, J., Khwaja, A., Vishwanath, T., & Zajonc, T. (2002). *Test feasibility survey Pakistan: Education sector.* Cambridge, MA: Harvard Kennedy School Working Paper.

Brown, H.G. (2004). *Language assessment: Principles and classroom practices.* New York: Pearson Education.

Fulcher, G. (1997). An English language placement test: issues in reliability and validity. *Language testing, 14*(2), 113-139.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London and New York: Routledge.

Giraldo, F. Validity and Classroom Language Testing: A Practical Approach1 *La validez y la evaluación de lenguas en el aula de idiomas: un enfoque práctico.*

Gronlund, N.E. (1998). *Assessment of student achievement. Sixth edition*. Boston: Allyn and Bacon.

Hashemi, A., & Daneshfar, S. (2018). A review of the IELTS test: Focus on validity, reliability, and washback. *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, *3*(1), 39-52.

Henning, G. (1987). *A guide to language testing: Development, evaluation and research*. Cambridge, Mass.

Huang, B. H., & Flores, B. B. (2018). The English language proficiency assessment for the 21st century (ELPA21). *Language Assessment Quarterly*, *15*(4), 433-442.

Hughes, A., & Hughes, J. (2020). *Testing for language teachers. Third edition.* New York: Cambridge University Press.

Im, G. H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9(1), 1-26.*

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–535

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*, 31–35. https:// doi.org/10.1111/j.1745-3992.2002.tb00083.x

Leung, C. (2022). Language proficiency: from description to prescription and back?. *Educational Linguistics*, *1*(1), 56-81.

Linn, R. L. (1997). Evaluating the Validity of Assessments: The Consequences of Use. *Educational Measurement: Issues and Practice*, *16*(2), 14-16.

Malone, M.E., & Montee, M. (2014). Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability. *TOEFL iBT ® Research Report*, 14-42.

Orozco, R. A. Z., & Shin, S. Y. (2019). Developing and validating an English proficiency test. *MEXTESOL Journal*, *43*(3), 1-11.

Shepard, L.A. (1993). Chapter 9: Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, *36*(8), 477-481.

Winke, P. (2011). Evaluating the validity of a high-stakes ESL test: Why teachers' perceptions matter. *Tesol Quarterly*, *45*(4), 628-660.

Wolf, M. K., Farnsworth, T., & Herman, J. (2008). Validity issues in assessing English language learners' language proficiency. *Educational Assessment*, *13*(2-3), 80-107.