JOURNAL OF FOREIGN LANGUAGE RESEARCH
PRINT ISSN: 2588-4123 ONLINE ISSN: 2588-7521
ww.Jflr.ut.ac.ir

Validation of a CEFR Based Test and Fairness in Assessment

Sayyed Mohamad Alavi *
Professor of TEFL University of Tehran
Email: smalavi@ut.ac.ir

Hossein Karami
Assistant Professor of TEFL University of Tehran
Email: hkarami@ut.ac.ir

Davood Sepahi
(corresponding author)
Ph.D. Candidate University of Tehran Kish International Campus
Email: dsepahi@ut.ac.ir

ABSTRACT
Fairness in assessment can be one of the important goals in the field of language teaching and assessment. Assessment experts acknowledge the relationship between validity and fairness. Some believe that validity itself is part of fairness. On the other hand, some believe that achieving fairness is in fact the achieving a test validity. The Common European Framework of Reference (CEFR) has developed a system of descriptors to achieve educational equality and assessment of learning. The purpose of this research is to validate a test consisting of 25 items based on CEFR descriptors. In order to obtain the necessary information to determine the validity of the test, 150 male and female participants answered multiple-choice questions. Then, the data were analyzed to determine the differential functioning of each item (DIF) by item response theory (IRT). In addition, a 75-item questionnaire was used to separate groups of participants based on their interests and goals. The results showed that in addition to the usual criteria for categorizing participants, including gender and linguistic and cultural background, personal interests and goals can be considered. It also became clear that these criteria could be used for determining test validity.

ARTICLE INFO
Article history:
Received: December 22, 2018
Accepted: July 14, 2020
Available online: Spring2022

Keywords:
"Fairness", "CEFR", "Assessment", "Differential Item Functioning", "Item Response Theory"

DOI: 10.22059/JFLR.2020.272015.588

Alavi, D., Karami, H., Sepahi, D. (2022). CEFR based test Validation and Fairness Assessment. Foreign Language Research Journal, 12 (1), 111-131.

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۲، شماره ۱، بهار ۱۴۰۱، از صفحه ۱۱۱ تا ۱۳۱

* Sayyed Mohamad Alavi Professor University of Tehran
** Hossein Karami Assistant Professor University of Tehran
*** Davood Sepahi Ph.D. Candidate University of Tehran

JOURNAL OF FOREIGN LANGUAGE RESEARCH
PRINT ISSN: 2588-4123 ONLINE ISSN: 2588-7521
ww.Jflr.ut.ac.ir

Validation of a CEFR Based Test and Fairness in Assessment

Sayyed Mohamad Alavi *
Professor of TEFL University of Tehran
Email: smalavi@ut.ac.ir

Hossein Karami **
Assistant Professor of TEFL University of Tehran
Email: hkarami@ut.ac.ir

Davood Sepahi ***
(corresponding author)
Ph.D. Candidate University of Tehran Kish International Campus
Email: dsepahi@ut.ac.ir

ABSTRACT
Fairness in assessment can be one of the important goals in the field of language teaching and assessment. Assessment experts acknowledge the relationship between validity and fairness. Some believe that validity itself is part of fairness. On the other hand, some believe that achieving fairness is in fact the achieving a test validity. The Common European Framework of Reference (CEFR) has developed a system of descriptors to achieve educational equality and assessment of learning. The purpose of this research is to validate a test consisting of 25 items based on CEFR descriptors. In order to obtain the necessary information to determine the validity of the test, 150 male and female participants answered multiple-choice questions. Then, the data were analyzed to determine the differential functioning of each item (DIF) by item response theory (IRT). In addition, a 75-item questionnaire was used to separate groups of participants based on their interests and goals. The results showed that in addition to the usual criteria for categorizing participants, including gender and linguistic and cultural background, personal interests and goals can be considered. It also became clear that these criteria could be used for determining test validity.

ARTICLE INFO
Article history:
Received: December 22, 2018
Accepted: July 14, 2020
Available online: Spring2022

Keywords:
"Fairness", "CEFR", "Assessment", "Differential Item Functioning", "Item Response Theory"

DOI: 10.22059/JFLR.2020.272015.588

Alavi, D., Karami, H., Sepahi, D. (2022). CEFR based test Validation and Fairness Assessment. Foreign Language Research Journal, 12 (1), 111-131.

* Sayyed Mohamad Alavi Professor University of Tehran
** Hossein Karami Assistant Professor University of Tehran
*** Davood Sepahi Ph.D. Candidate University of Tehran

1. Introduction:

Fairness in assessment is one of the most important goals in the field of language teaching and testing, as it has attracted the attention of scientists in recent decades (McNamara 2005, Kunnan 2000 & 2004, Xi, 2010). Yet, some do not consider these issues to be very important (Davis 2013). However, research is critical to defining and determining the limits of fairness as well as its relationship to test validity. McNamara and Ryan (2011) define equality: "Ensuring the quality of the test, especially its psychological quality, and equality in procedures for the individual and subgroup of test participants and the adequacy of representation of the structure in materials and methods of experiments (p. 163)". This definition is beyond the definition of justice in the test. Identifying the starting point, the type of training materials, the quality of the assessment, as well as the educational objectives are among the information that are important for creating fairness in assessment. Research on fairness has been done in both theoretical and practical ways. Each, in turn, transforms the perspective of stakeholders.

The Common European Framework of Reference (CEFR), developed with extensive research support, can provide a basis for achieving fairness in evaluation. This framework was introduced in 2001 (Council of Europe, 2001) and is recommended to be used gradually in the education system of EU member states. In addition to these countries, research on the use of CEFR at various levels has been carried out in Australia (McNamara and Ryan, 2011), China (Huang & Jia, 2012), Japan (Nagai & Odevier, 2011) Taiwan (Wu & Wu, 2007) , Turkey (Peachy, 2012), Canada (Elatia, 2011). Also, aligning well-known international tests such as TOEFL and IELTS with the classification of language proficiency levels according to CEFR criteria is significant. In this way, the scope of application of CEFR has been doubled with these well-known tests, so that learners can measure the level of their learning in relation to these tests. Also, well-known international publishers have defined and aligned their textbooks with CEFR, which again adds to the scope of application of this framework.

The use of CEFR is not only limited to the field of language teaching, but also to the preparation of tests for English-speaking students in non-language fields. In a study, Shaw and Imam (2013) showed that CEFR can be used to test the level of cognitive proficiency of candidates' academic language. In their view, the minimum language skill required for candidates is B2 level. Also, the skill level of C1 can be a good ground for progress in learning subjects such as history and geography. At their suggestion, CEFR could be a platform for providing a language package suitable for non-linguistic educational content. That is, teachers who teach non-linguistic content

- such as history or geography - can use this language package to better teach educational content to non-English speaking students.

On the other hand, the use of CEFR has so far had a profound effect on language assessment (Little, 2014). According to Little, language learning and evaluation are closely linked. In this view, the learner is responsible for learning. Personal supervision of learners' learning can be done using the "can do" tables. Research has been done in this area that confirms Little's opinion.

CEFR features

The documents related to CEFR state that the description of language situations is defined based on the needs of learners regardless of the first and second languages. Each of these definitions can be used for learning, teaching, and assessment. The groups involved in language teaching are: policy makers, content developers, teachers, and other stakeholders each of whom can the CEFR descriptors optimally and coherently. For example, the preparation of teaching materials requires a precise definition of a possible linguistic situation or scenario. There are examples of these situations in the CEFR published documents. Learners should be able to clearly help with their learning goals and be able to assess their success or failure at any stage of learning, i.e. self-assessment. When test makers are aware of the goals and definitions of multiple learning situations and levels, they can prepare tests tailored to the goals of each stakeholder. Thus, the validity of the test score interpretation is further guaranteed (Council of Europe, 2001, p. 5).

CEFR descriptors:

These descriptors are designed to recognize people's language performance in real communication situations. The initial triple classification of these descriptors is from top to bottom, respectively: proficient user (C1-C2), independent user (B1-B2) and basic user (A1-A2). These levels of language proficiency with clear sentences in understanding: reading comprehension, understanding: listening comprehension, speaking and spoken interaction, as well as writing are described. An example of the two described levels A1 and B2 can be seen in Table 1. These descriptors are also known as "can do statements."

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 12, Number 1, Spring 2022, Page 111 to 131

Council of Europe

Levels of the Common European Framework of Reference (CEF*R*)

	UNDERSTANDING Listening	UNDERSTANDING Reading	SPEAKING Spoken interaction	SPEAKING Spoken Production	WRITING Writing
A1	I can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.	I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.	I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I'm trying to say. I can ask and answer simple questions in areas of immediate need or on very familiar topics.	I can use simple phrases and sentences to describe where I live and people I know.	I can write a short, simple postcard, for□example sending holiday greetings. I□can fill in forms with personal details, for example entering my name, nationality and address on a hotel registration form.
B2	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.	I can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. I□can understand contemporary literary prose.	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	I can write clear, detailed text on a wide range of subjects related to my inter-ests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters

Table 1. Descriptors A1 and B2 in CEFR Language Skills (Council of Europe, 2001) □

CEFR limitations:

This framework was originally developed for adults. For this reason, descriptors need to be reviewed in order to provide educational materials as well as special tests for children and adolescents. Studies show that levels defined in CEFR are beyond children's cognitive comprehension. The transition from one level to another is not in line with children's cognitive development. In order to address these shortcomings, the following objectives should be considered: Definition of as many identifiers as possible that represent the world and language of children, do not define a level higher than B1, yet levels in between (e.g. A1 +, or A2 +) should be defined so that children experience real progress. One of the drawbacks of CEFR (Alderson 2007) is the lack of empirical evidence in the validity of its application.

2. Research background:

CEFR alignment

Internal alignment

Aligning CEFR with existing educational and teaching systems is one of the effective methods in applying and researching CEFR. These studies have been done in second language teaching and evaluation systems in each country with different historical, cultural and educational background. This process demonstrates an attempt to establishing the validity of the application of

CEFR in a variety of contexts. Because, the necessity of covering issues arising from multilingual and multicultural phenomenon in the fields of education, learning and evaluation, has been discussed since the beginning of the emergence and design of CEFR. In a survey, Negishi et. al. (2012) analyzed the results of a vocabulary test in a large group of Japanese participants. The findings showed that the order in which the compound verbs were presented did not necessarily correspond to the propositions in the CEFR descriptors. However, they acknowledged that the Japanese language may have contributed to the situation, and suggested that similar research be conducted in other languages. In another study on the use of CEFR in the design of curricula based on the communication needs of language learners at the university level, Arsalan and Oznici (2017), after mentioning the benefits, suggest using this framework in the design of language teaching programs. Among these studies is the teaching of general English to students in non-language fields in Vietnam. It has been reported that alignment of CEFR in the education system has led to a qualitative change in students' self-assessment, end-of-semester examinations, and student interaction (Le H.T.T., 2018).

External alignment

CEFR has been used to prepare basic tests for major and non-European languages such as Arabic. In this field, descriptors of practical language skills are used to define the level of familiarity and mastery of the language, and based on these definitions, the skill assessment test in that language is obtained. For example, Ali and colleagues (2018) states in a report on a part of the TALP: Test of Arabic Language Proficiency: "Many points have been edited to match the requirements of the Arabic language, therefore, the descriptive framework of this research contains the following points:

• Difficulty of words at each level and how to select and sort

• Grammatical elements, classifications and structural sections, as well as processes and relationships

• Vocabulary elements and processes

• Semantic relationships

• Recognition and distribution skills at different levels

• Social groups in the destination community that learners should be familiar with

In addition to the above, descriptions of the following issues are also provided:

• Principles of selection of micro and macro roles and their stratification

• Identify areas and why they are important

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 12, Number 1, Spring 2022, Page 111 to 131

• Determining language activities that should be included in the test and curriculum?

• Determine the number of teaching hours according to the language levels required

CEFR Intercultural research

An important question about efficiency, and especially fair validity of CEFR, is very important in different cultural areas. Can CEFR, which has been introduced as a macro-reference in multicultural linguistic planning, be applied to obtain the desired validity in planning at a micro level? In an intercultural study, the content of third- and fourth-grade English language textbooks in primary schools in Turkey and Portugal was compared with descriptive references to CEFR at level A1. The results of the comparison showed differences and similarities in the performance of CEFR in each of these two countries. Suggestions based on the research results include teacher training in familiarity with the characteristics of CEFR, preparation of multicultural content appropriate to the age of learners at the elementary level, and finally advice on continuity of curriculum planning in the cultural and linguistic framework of both countries (Guerra, 2018).

The use of CEFR in leveling the first language

An interesting example is the use of CEFR to design the first language test - not the second language. In the research, general descriptive requirements for learning elementary level and structure, as well as assessment and measurement of Tatar language at level A 2 have been used by the system of CEFR descriptors. According to researchers, the definition of Tatar language in this system facilitates the transfer of new methods of teaching a second language. In fact, the CEFR is used as a bridge to exchange methods of teaching, assessment and evaluation. The researchers wanted to go beyond Russia's internal borders and establish an international connection and make the Tatar language known to other cultures. Also, the needs assessment of the Tatar language will be coordinated with an advanced field, so that as a result of this connection, it will lead to the continuous updating of educational methods in the field of Tatar language teaching and learning (Shakerova, 2018).

Citizenship Language Assessment

In immigrant destination and multilingual countries - Switzerland - and multicultural - United States - the linguistic knowledge and ability of immigrants and compliance with the linguistic norms of the destination community has always been a complex issue. In Switzerland, for example, the process of accepting immigrants must take place in such a way that the understanding of culture and language in society is accomplished as much as possible. Therefore, in order to standardize all processes, CEFR descriptors were translated

to comply with the minimum and maximum requirements of the laws of the cantons (Arrighi, and Piccoli, (2018).

CEFR based language corpus

Accumulation of texts from students' writing of texts (second language) is one of the appropriate strategies for constructing a linguistic corpus. This corpus can be used extensively in preparing educational materials and necessary tests. To this end, in Sweden, which is an immigrant country, texts written by Swedish language learners in various language learning systems have been prepared based on CEFR descriptors. The purpose of forming this language corpus is to critically redefine and evaluate the validity of teaching methods and measuring the current system and to review the content and provide teaching and evaluation with equal status for language learners (Megyesi, et al., 2018).

Materials preparation

In compiling textbooks, several points such as linguistics, sociology and pragmatics are very important. Advertising, for example, is just one of the sources for writing texts to familiarize learners with a body of linguistic information in the target language, for which purpose CEFR can be used to categorize the language levels (Pérez de la Calle, 2018).

Levels of ability: listening and speaking

In order to determine the level, Chama has effective applications in all four language skills. For example, to determine the complexity of accuracy and fluency of speaking CEFR levels is used. In a study Karami and colleagues selected texts which were considered in the range of A1 to C1 to determine the full coverage of participants' speaking ability (Karami et al., 2018). In another study (Domna, and Zafiri, 2018) to test the learning rate of private school language learners, A1 level tests were given to participants as a pretest. The aim of this study was to determine the difference between a new and the conventional teaching in private schools. For this purpose, tests relevant to CEFR levels were used, because according to the researchers, the educational program of these schools has been prepared and presented based on that framework.

Evaluation of progress

To assess students' progress at higher levels, common test can not be used. Since, these tests do not indicate linguistic particulars such as collocation at high levels B2 and C1. To determine this level of skill in learning French, a number of noun-verb collocations were prepared. It was found that the precise use of collocations indicates a high level of linguistic knowledge (Lundel et al., 2018).

Theoretical definition of validity

In assessment (and educational evaluation), topics are related to ontology, epistemology, and ethics. What we know as truth, how to get to it, and the consequences of the

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 12, Number 1, Spring 2022, Page 111 to 131

assessment (Fulcher, quoted from Hondrich, 1995. p. 666), almost all in assessment and educational evaluation agree on the principles of ethics, fairness and consequences in people's lives. However, the differences of opinion on how to achieve a single definition and how to achieve operational definitions resulting from the definition of concepts are mostly related to ontological and epistemological methods (Razavipur, K., 2019).

The issues raised in the work of further cognition end in two perspectives. In one view, there is "truth" independent of us. In another view, truth is what we perceive from truth. For each of these perspectives, there are several arguments throughout the history of science that we will not address here. Rather, we express the impact of each of these two perspectives on educational assessment and evaluation.

Because the discussion of the nature as well as how the results of the test and evaluation affect outside the scope of statistical calculations, and on the other hand the relationship between statistical findings and interpretation with interpretations resulting from the test score in determining the validity it has an effect, it needs an argumentative method to give a clear explanation of the statistical and other data obtained from the test. In the following, while reviewing the definition of validity, the argumentative method about the relationship between test results and its consequences, as well as its interpretation, is introduced.

The simplest common definition of validity is: "A test measures what it is intended to measure." This definition seems very simple and understandable in the first reading. But the issues involved in teaching, assessment, and evaluation are so complex that this definition cannot be enlightening. Therefore, assessment and evaluation specialists seek to find a definition so that in addition to interpretation of test use and its validity, they can investigate the effect of the score on the person's life and subsequent consequences. By presenting the theory of "validity as a unitary concept", Messick put validity concepts into a single pattern. Previously, a multi-part validity was defined, each covering part of the reality of the test and its consequences. Another definition that identifies validity as one of the basic components of tests is as follows: the extent to which the interpretation and application of test results are validated by users with theory and evidence (American Psychological Association, 1999; p. 9, cited in Cummins, 2012).

Argument base validity

This method was developed by Kane (1992-2006) to justify the reasoning of language tests following the invention of the Toulmin argumentative method (2003). This idea consists of two parts: first, the expression of the interpretive argument, then the compilation and evaluation of the evidence

of the argument, and also considering the potential cases of potential counterevidence of the argument. In this regard, Bachman (2005) emphasized the need for research and application of argumentative methods to validate a test. Bachman and Palmer (2010) formulated conceptual relationships between test-based interpretation, and decision-making based on results, and consider the consequences, as well as attention to fairness in evaluation within an assessment use argument. Kunnan (2004) presents two complementary frameworks in the conceptualization of fairness, one is the framework of test fairness and the other is the framework of social context. The test framework includes validity, lack of bias, access, test execution, and social consequences, and each of these qualities is very important in developing and applying a test. Another framework monitors environmental factors affecting the fairness of the test (p. 27). Since it is not easy to define and calculate the impact of each of these factors, at least these major factors affecting - almost - all environments when interpreting test results should be considered. These major factors are: political, economic, educational, social, cultural, and legal and ethical (Kunnan, 2008, p. 240).

Fairness in assessment

Fairness in a test is related to the consequences of the test results for individuals and groups or society. This concept is related both to validity and to some extent as an indicator of the measurement of language ability and thus to social equality (Davis et al., 1999). Similarly, it can be said that research on fairness is directly related to the test and the application of its result in the social context. On the other hand, the lack of bias in a test, in the psychometric approach, can provide the concept or quality of fairness. The question of the existence of this quality in the test actually arises when different groups of test takers are in the assessment process. Does the test only measure language proficiency and does not lean in favor any of the participants? To answer this question, various statistical methods have been used and various groupings such as linguistic, gender, and cultural have been studied more. This research starts from the beginning of the process of preparing items to administration and scoring method and even continues in the type of interpretation of social results and consequences. Test developers and test users use appropriate statistical methods to reduce the unfair effects on the test to optimize the test as much as possible (Kunnan, 2004; McNamara and Rover, 2006; Bachman and Palmer, 2010). The process of examining the effect of fairness in a test can be seen in the works of researchers such as Messick (1989), Kane (2000), Kunnan (2004) and Xi (2010).

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 12, Number 1, Spring 2022, Page 111 to 131

This line of research shows that fairness in a test is related to equality, validity and lack of bias. In this regard, according to Xi (2010), fairness is the same validity that is defined equally for groups and is applied in the preparation, implementation and interpretation of test results. In order to investigate the relationship between fairness and validity, according to Kunnan (2000), validity is defined by the construct definition of an equal structure to interpret the test result for all groups of subjects. For him, this is possible by controlling the content bias the way the item is presented, item functioning, and the language of the tests. The validity and fairness in Kane's (2010) perspective are very close to each other, and the degree of over-coverage and under-coverage of each varies according to their definition in different situations. However, in his opinion, in general, neither of these two includes the other, and each has an independent identity.

McNamara and Rover (2006) without providing a clear theoretical definition and a definite solution claim that the issues related to fairness and the social role of the test are not sufficiently codified. They point to the need for social fairness and the prevention of test bias in favor of or against social groups. However, they suggest that research on the individual differences of test takers and its effect on their performance in the test and considering the criteria of professional ethics is possible through the differential item functioning (DIF).

Significance of the Study

CEFR is gradually playing a key role in issues related to language teaching and assessment among European countries (e.g. Negishi et al. 2012). Numerous researches are carried out in different branches such as education, assessment, preparation and compilation of educational materials, teacher training, etc. based on CEFR. This trend in research is not limited to European countries and is ongoing in other parts of the world. This highlights the importance of the role of CEFR in the future of language teaching - and related matters. Therefore, recognizing this framework and researching how to use or even criticize its achievements is inevitable.

Creating a connection with a test based on CEFR requires purposeful planning on the one hand, and then administration and calculating the data within a statistical framework. On the other hand, linking the results of the test with the contextual factors and purposes of the test takers and finally interpreting the results of the test to explain and justify the validity of the test. Virtually, the extrapolation of test results and related interpretations is performed.

Adapting the test score to common standards (e.g. CEFR and ACTFL) is usually expressed in specific words. The concept attributed to these sentences, each

directly related to the definition of language learning levels and perceptions of stakeholders may vary from context to context. In addition, people with different goals need a clear understanding of the concept of test results and their relationship to their personal purpose. Providing evidence for the relationship between test results and related interpretations is an important part of the validity argument of the results (Bachman, 2003). Therefore, determining the framework of the basic process in validity argument is the result of a test that has been created for a specific purpose with linguistic descriptors. In this regard, several issues should be considered. Who is interested in the validity process of the test and the meaning of the results for each group of stakeholders should be expressed in clear language and according to the goals and objectives of each group. The other is that the test validation process does not end in one step, but the results give a different interpretation depending on the change in the circumstances and the beneficiaries and the stated goals.

Test results provide information to test developers and users to use as evidence for validity argument, which in turn influences the connection between results and interpretation and the corresponding decisions (Kane, 1992; Xi, 2008). The choice of statistical analysis and interpretation of the results depends on whether the test is norm referenced or criterion referenced. In this study, the test is assumed to be of the criterion referenced type because it is based on the framework of the CEFR. One of the features of the present study is that it establishes a relationship between the method of validity argument and the application of its test with the frameworks presented in CEFR in determining the level of test takers based on their cultural interests. In other words, the mental conditions of the participants are involved in interpreting the results through a questionnaire. In this way, the information obtained from statistical analysis provides information to decision makers and other stakeholders that have as much effect as possible on the validity of the interpretation and its consequences on the lives of the test takers.

3. Research Method:

Operational definition: Various statistical and computational methods have been used in operational definition. One of these methods is the differential item functioning. To calculate DIF, participants are usually divided into target and reference groups based on some criteria. This division varies according to the issues facing the community and the research priorities. In societies such as the United States, the issue of the racial background of individuals (Spanish and other minority languages) is considered in the study. Also, other issues such as gender are a complex and worthy of research in most societies. Here, it should be

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 12, Number 1, Spring 2022, Page 111 to 131

noted that these divisions should not be based on the process of previous research, but on their cognitive perspective and what they should be according to the immediate need. For example, differentiated groups can be defined based on motivation and purpose or position and access to educational factors. The statistical methods used are each tailored to different situations and definitions.

In the present study, the participants' answers to the questions of the English language proficiency test, which are based on CEFR descriptors, were corrected and categorized. Test items were identified and categorized in terms of language proficiency construct. Then, the obtained results were analyzed in item response theory (WINSTEPS Version 3.92).

Statistical population

The statistical population of this study is 150 students - male and female - in the fields of petroleum and petrochemical engineering, fluid and solid mechanics, computer and information, electricity and chemistry. They have been admitted to these courses by taking the entrance exam. Their age ranges from 18 to 23 years. Most of them took part in a general English language course in an academic year.

Research instruments

In this study, a 25-item English language proficiency test was used to determine the level of English language proficiency of the participants. The selection of items is based on the linguistic and skillful descriptors of CEFR. Questions are ordered from easy to difficult. No negative score was considered for incorrect responses. In a classification of constructs underlying the test items, five experts were asked to determine the constructs of each item. Table 4.1 summarizes the opinions of the experts:

Table 2 Summary of experts' views on constructs of items,

(2) collocation and vocabulary (3) Grammar (4) Pragmatics

1	Expert1	Expert2	Expert3	Expert4
2	6, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19, 22, 23, 24, 25	6, 8, 9, 10, 11, 12, 17, 18, 19, 21, 22, 23, 24, 25	6, 8, 9, 10, 11, 12, 14, 15, 17, 18, 19, 22, 23, 24, 25	3, 6, 8, 9, 10, 11, 12, 17, 18, 19, 21, 23, 24, 25
3	1, 2, 3, 4, 7, 13, 16, 20, 21	7, 13, 14, 15, 16, 20	2, 3, 7, 13, 16, 20, 21	7, 13, 14, 15, 16, 20
4	4, 5	1, 2, 3, 4, 5	1, 4, 5	1, 2, 3, 4, 5

In addition to the above instruments, the Persian version of the 75-item questionnaire (Dörnyei, & Taguchi, 2009) was used to ask participants' 'views on their future educational opportunities, future occupational opportunities', and their 'cultural views on the target language community'. Then these categories were used as the basis for groupings needed to perform differential item functioning and other necessary analyzes.

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۲، شماره ۱، بهار ۱۴۰۱، از صفحه ۱۱۱ تا ۱۳۱

To test the relevance between the underlying constructs of the items and the items in the questionnaire groupings Pearson correlations were performed. For the above mentioned groups the same items acted differently as shown in the table below:

Results of IRT statistical analysis

Pearson correlation was used to test the relationship between underlying constructs and test group items the items behaved differently as shown in Table 4.2.

Table 3

Relationship between CEFR test components as indicators of AUA (N=150)

		Vocabulary Collocation	Grammar	Pragmatic competence
Future career opportunity	R	.053	.111	-.012
	Sig	.519	.178	.887
Future Education opportunity	R	-.006	.018	-.062
	Sig	.945	.827	.450
Cultural Attitude	R	-.065	.025	-.020
	Sig	.426	.762	.808

A study conducted by Nematzadeh (2018) shows that cultural characteristics and specific backgrounds such as age, gender, and mother tongue did not differ in the test. In a study using the differential item functioning (DIF) on gender in the High Stake Language Ability Test (NUEEFL), Bordbar (2020) concludes that scores from the test "construct irrelevant variance are not structural and the overall equality of the test is not confirmed."

The effect of CEFR test scores on Assessment Use Argument

All items were categorized into three general groups: vocabulary, grammar, and pragmatic competence. Then, a linear regression application was performed to test the relationship between these three groups of CEFR based items Assessment Use Argument. The results in the table show that grammar items have more predictive power than the other two categories of items. That is, items 1, 3, 7, 13, 16, 20, and 21 provide more information about assessing the language proficiency of test participants. Interestingly, the same items are of particular importance in analyzing the differential item functioning, which will be presented later. Another noteworthy point is that items 1 and 4 of the pragmatic competence group in comparison with the

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 12, Number 1, Spring 2022, Page 111 to 131

other two parts of grammar and vocabulary and collocation showed more predictive power for the category of cultural attitudes of the questionnaire. This information is summarized in the tables below.

	Vocabulary Collocation	.003	.121
	Grammar	.200	.166
	Pragmatic competence	-.074	.154

a. Dependent Variable: Future career opportunity

Future career opportunity = 72.20 + .003 (Vocabulary collocation)

Table 4. Coefficients[a]

+.200 (Grammar)

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	T	Sig.
1	(Constant)	72.200	8.225		8.779	.000

.074 (Pragmatic competence)

Table 5 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	T	Sig.
1	(Constant)	69.964	7.110		9.841	.000
	Vocabulary Collocation	.005	.105	.006	.052	.959
	Grammar	.048	.144	.034	.334	.739
	Pragmatic competence	-.107	.133	-.073	-.804	.422
a. Dependent Variable: Future Education opportunity						

Future education opportunity = 69.96 + .005 (Vocabulary collocation)

-.107 (Pragmatic competence)

+..048 (Grammar)

Table 6 Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	47.835	7.819		6.118	.000

	Vocabulary Collocation	-.131	.115	-.121	-1.137	.257
	Grammar	.146	.158	.093	.923	.357
	Pragmatic competence	.011	.146	.007	.074	.941
a. Dependent Variable: Cultural Attitude						

Cultural Attitude = 47.83 -.131 (Vocabulary collocation)

.146 (Grammar)

.011 (Pragmatic competence)

DIF analysis of CEFR based on indicators of assessment use argument

A DIF analysis was used to estimate the rate of CEFR based test bias relative to the ratios of the evaluation application. Most of DIF studies have focused on finding items that have gender, racial, and so on bias. The idea of classification in the present study is not based on the natural characteristics of the participants. Because they do not anything to do in the selection of these features and they are simply acquired. While they possess traits throughout their lives, although they are influenced by natural traits, individuals themselves contribute to the selection, development of these secondary traits. In this way, it can be said that these characteristics are based on the choice and decision in their lifestyle and their perceptions and intentions.

In addition, it can be said that to examine the effect of the test on the performance of the participants and the consequences of the interpretation results obtained from the test scores it is directly related to the validity of the test. Their career and educational life, as well as their cultural attitudes, are influenced by the secondary characteristics of their choice. Also, job selections in the present age can be based on people's interests and choices. This diminishes gender and language and other characteristics of this type in our national society.

Based on this, it can be said that those natural features that are a problem in multilingual and multicultural environments such as the United States are not very important for us. Rather, such issues should not be interfered with in studies of fairness or non-bias because of the same mother tongue and race. Table 4.5 shows information about participants' performance and questionnaire.

Table 7

Descriptive Statistics for the subgroups (N=150)

	Mean	Std. Deviation

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 12, Number 1, Spring 2022, Page 111 to 131

Future career opportunity	79.58	15.12
Future Education opportunity	68.33	13.01
Cultural Attitude	48.45	14.35

Since the standard error deviation of all items in both groups is above 0.000, some DIF in the data is observed. DIF measure is the difficulty of an item for a group, when all other influential conditions are the same. The DIF contrast is a "measure of effect" based on Logit, and the difference between the two DIF criteria, the DIF size, is classified into two groups. Positive DIF contrast indicates that the item is more difficult for the comparing group. By definition, one group of participants receives a higher score than the other. The reason for this difference, according to Linacer (2019), could be due to the following.

1. The group is naturally able to act on the item, the other group is better than normal.

2. The group is naturally able to act on the item, the other group is worse off than normal.

3. An item has its normal difficulty for one group, but is more difficult for another group than normal.

4. The item has its usual difficulty for one group, but is easier for the other group than normal.

It is important to note that statistical information does not differentiate between these justifications, but it is the relationship between the results of DIF analysis and the necessary decisions based on specific goals that makes the differences meaningful. So if the items are in favor of one group and at the same time to the detriment of another group, we say that the test consisting of these items is unfair. From the beginning of determining the purpose for preparing the test and while writing the items and then in the initial and main administration and also when scoring and interpreting the scores, statistical data should be combined with qualitative information and in order to achieve fairness in the test conditions. However, the following is statistically significant for DIF results in terms of Linacer (2019):

1. The size of the impact of DIF

2. Group classification size

The results of DIF analysis in tables and graphs show information about each of the items that have had a significant amount of group bias.

DIF future career opportunities Explained

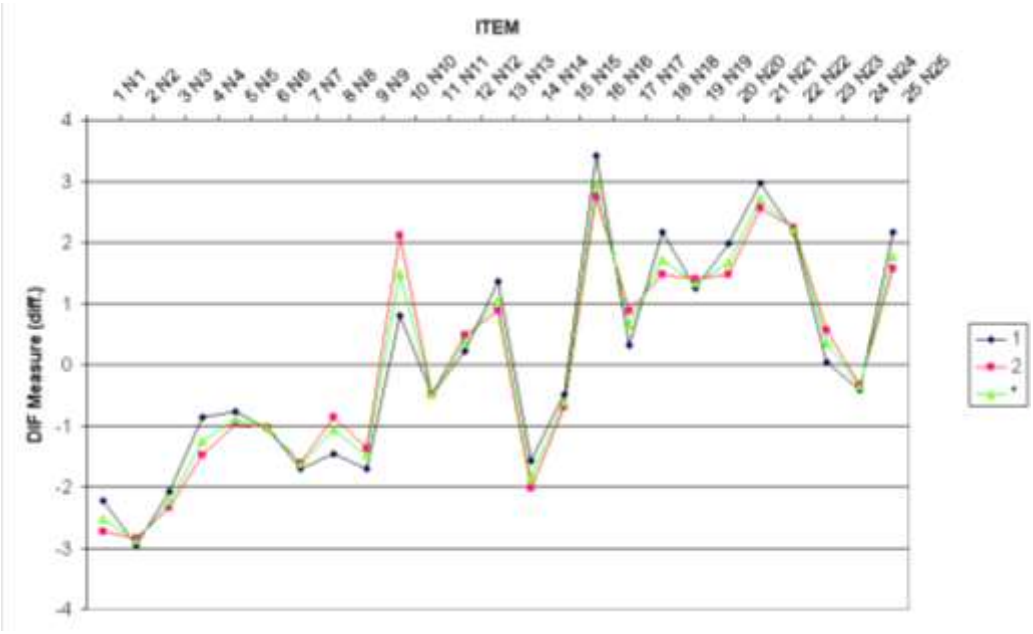

Figure 1 future career opportunity DIF measure explained

As can be seen in Figure 1, items 1 to 8 have almost identical DIF for both groups of respondents. However, items 9 and 10 are somewhat more DIF than previous items, and there is little difference between groups 1 and 2. However, items 11 and 12 are not as distinct as the previous items, and at a lower level in groups 1 and 2 have almost the same function. Item 14, but it is noteworthy that its resolution differs from previous items and has many similarities in both groups 1 and 2. Item 15 is three levels above the reference line (0) and has the same properties for both groups. Another interesting point is that item 24 and item 10 are on the same level. However, the ups and downs of items as a whole show that participants feel an upward trend in interaction with items.

Table 8 DIF for the future career opportunity group

Item Number	Name	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
							T	df	prob.
4	N4	1	-.86	2	-1.48	.62	1.57	119	.1183
8	N8	1	-1.46	2	-.86	-.61	-1.50	106	.1356
10	N10	1	.80	2	2.12	-1.32	-2.74	144	.0070
16	N16	1	3.42	2	2.75	.67	.78	95	.4372
18	N18	1	2.17	2	1.48	.68	1.26	101	.2106

Items 4, 16 and 18 are simple in favor of group two and items 8, 10 in favor of group one are simple.

DIF explanation of future education opportunities

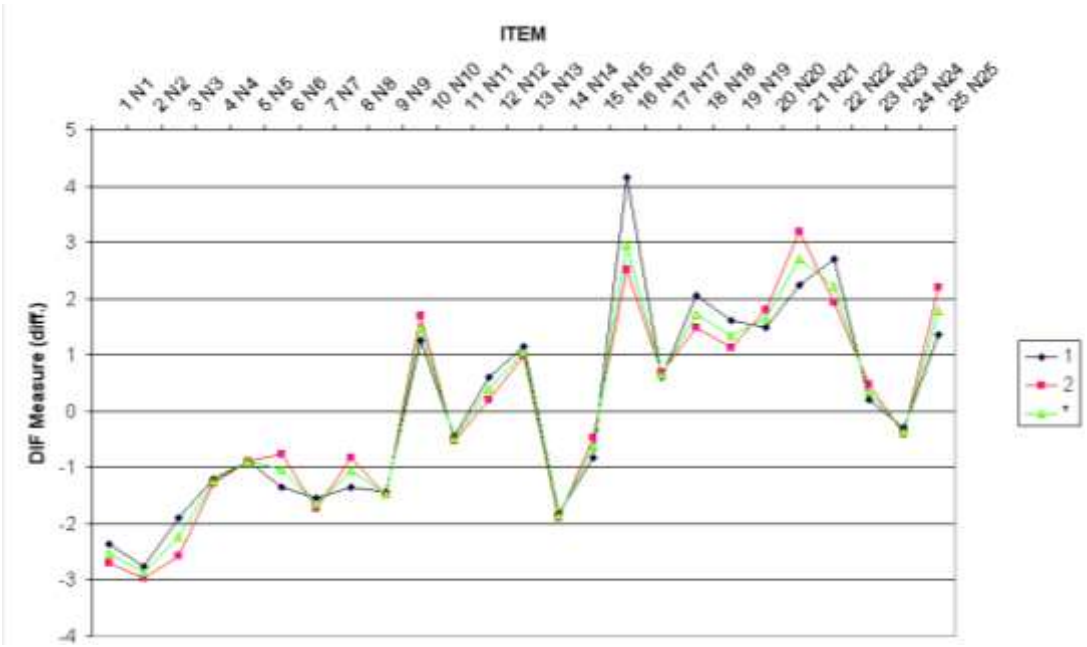

Figure 2 DIF measure of future education opportunities

Figure 2 shows DIF in terms of the difference of responses according to the future education opportunities of the participants. Two items 10 and 20 each show a greater distinction for group 2 participants. Also, items 19 and 25 are secondary to distinction. Item 18 shows a higher distinction for group 1.After that, item 3 shows more distinction for group 1. In the general view, there is not much similarity between the diagram and the size of the DIF distinction, as well as the future career opportunities of the participants. Perhaps this discrepancy is related to other features that should be considered in another analysis.

Table 9 DIF for future education opportunities group

Item Number	Name	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
							t	df	prob.
3	N3	1	-1.91	2	-2.57	.67	1.43	147	.1538
16	N16	1	4.17	2	2.52	1.65	1.50	88	.1367
21	N21	1	2.24	2	3.19	-.95	-1.37	146	.1729
22	N22	1	2.70	2	1.93	.77	1.21	118	.2279
25	N25	1	1.37	2	2.20	-.83	-1.63	147	.1050

پژوهش‌های زبان‌شناختی در زبان‌های خارجی، دوره ۱۲، شماره ۱، بهار ۱۴۰۱، از صفحه ۱۱۱ تا ۱۳۱

Items 3, 16 and 22 are in favor of group 2 and items 21 and 25 are in favor of group one.

DIF explanation of cultural attitude

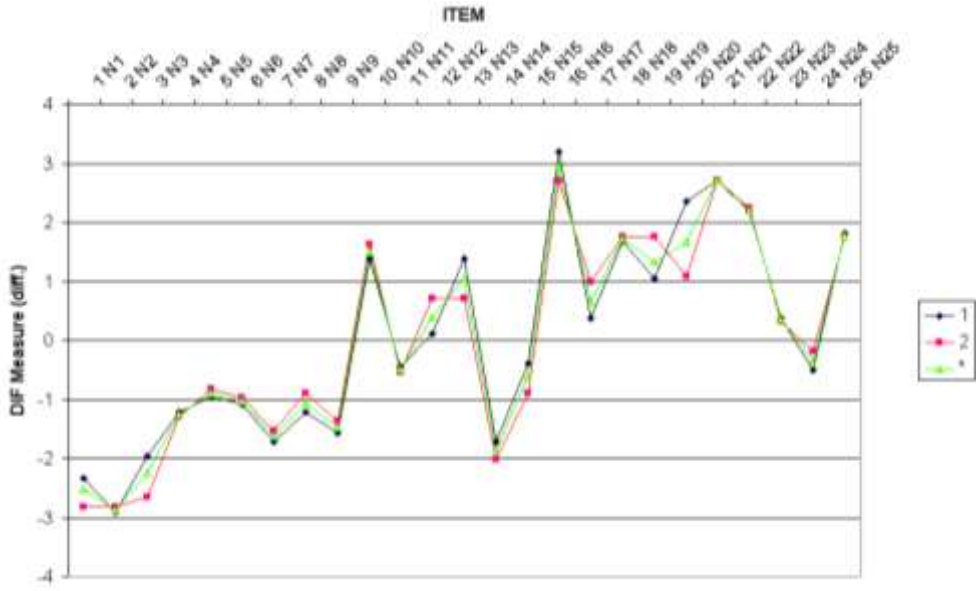

Figure 3 DIF measure for cultural attitudes

Table 10 DIF for the cultural attitude group

Item Number	Name	Person Class	DIF Measure	Person Class	DIF Measure	DIF Contrast	Rasch-Welch		
							t	df	prob.
3	N3	1	-1.96	2	-2.66	.69	1.42	129	.1579
12	N12	1	.12	2	.72	-.60	-1.56	140	.1206
13	N13	1	1.39	2	.72	.67	1.59	147	.1148
17	N17	1	.38	2	.98	-.60	-1.51	138	.1340
19	N19	1	1.04	2	1.76	-.72	-1.54	131	.1269
20	N20	1	2.35	2	1.08	1.27	2.51	144	.0133

Items 3, 13 and 20 are in favor of group 2 and items 12, 17 and 19 are in favor of group one.

7. Discussion

The best tests are valid only where they are defined and prepared. Therefore, for each application in another situation, they need to redefine the validity. Definition of validity is volatile. In other words, each test has validity for its defined purposes in a given time and place. CEFR provides a suitable context for researching and realizing the validity of the language test, because it is defined for all languages and does not depend on a specific language. The assessment use argument is a logical structure in that the data in a given and comparable framework are provided and the necessary evidence to determine the validity of each test. This framework can accommodate a variety of statistical calculations that have been prepared quantitatively and provide the necessary

JOURNAL OF FOREIGN LANGUAGE RESEARCH, Volume 12, Number 1, Spring 2022, Page 111 to 131

qualitative evaluation according to different conditions. Also, the data of the present study indicate that the differential item functioning of a test should not be considered fixed and definite, but it is necessary to interpret the use of the scores according to the specific conditions of the test, participants, their goals, and those of organizations and stakeholders.

References

Alderson, J.C., (2007). *The CEFR and the Need for More Research* The Modern Language Journal, 91(4), pp. 659-663.

Ali, Z. M., Ali, F., Radzuan, N. M., Alwi, N. N. M., Abu, N. L., & Kassim, Z. (2018). Contextualising the CEFR: the Universiti Malaysia Pahang English language proficiency writing test. In 11th Annual International Conference of Education, Research and Innovation (ICERI 2018) (pp. 4892-4902).

Arrighi, J. T., & Piccoli, L. (2018). SWISSCIT Index on Citizenship Law in Swiss Cantons: Conceptualisation, Measurement, Aggregation. Université de Neuchâtel.

Arsalan, A. & Özeinci, S. (2017) *A CEFR-based Curriculum Design for Tertiary Education Level* International Journal of Languages ' Education and Teaching. Volume 5, Issue 3, p. 12-36

Bachman, L. F. (2003). Constructing an assessment use argument and supporting claims about test taker-assessment task interactions in evidence-centered assessment design. MEASUREMENT-LAWRENCE ERLBAUM ASSOCIATES-, 1, 63-65.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly,2, 1–34*.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, England: Oxford University Press.

Bordbar, S. (2020). Differential Item Functioning of University Entrance Exam: Using Rasch Analysis. *Foreign Language Research Journal*, 10(1), 37-55. doi: 10.22059/jflr.2019.278170.611

Council of Europe. (2001). *Common European Framework of Reference for Languages*. Cambridge, UK: Cambridge University Press.

Cummins, A. (2012). *Validation of language assessments*. The encyclopedia of applied linguistics.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, England: UCLES/Cambridge University Press.

Davis, A. (2013). Fifty Years of Language Assessments. In The Companion to Language Assessment, A. J. Kunnan (Ed). doi:10.1002/9781118411360.wbcla127

Domna, A., & Zafiri, M. (2018). A Case Study of Two Groups of A1 Level Students in English. Education and Linguistic Research.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. Routledge.

Elatia, S. (2011). *Choosing language competence descriptors for language assessment: validity and fairness issues*. Synergies Europe, 6, 165-175.

Fulcher, G., & Davidson, F. (Eds.). (2013). *The Routledge handbook of language testing*. Routledge.

Guerra, L., Gonçalves, O., Fisne, F. N., & Gungor, M. N. (2018). A CEFR-based comparison of ELT curriculum and course books used in Turkish and Portuguese primary schools.

HUANG, T., & JIA, G. (2012). *The feasibility study of linking language test to the CEFR—Taking College English Test as an example* [J]. Foreign Language Testing and Teaching, 1, 007.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*, 527–35.

Karami, M., Jafarigohar, M., Tajeddin, Z., & Rouhi, A. (2018). Input-induced Variation in EFL Learners' Oral Production in Terms of Complexity, Accuracy, and Fluency. Iranian Journal of English for Academic Purposes, 6(2), 70-85.

Volodina, E. (2018, November). Learner Corpus Anonymization in the Age of GDPR: Insights from the Creation of a Learner Corpus of Swedish. In Proceedings of the 7th Workshop on NLP for Computer Assisted Language Learning (NLP4CALL 2018) at SLTC, Stockholm, 7th November 2018 (No. 152, pp. 47-56). Linköping University Electronic Press.

McNamara, T. (2005). 21st century shibboleth: Language tests, identity and intergroup conflict. Language Policy, 4(4), 351-370.https://doi.org/10.1007/s10993-005-2886-0

McNamara, T., & Ryan, K. (2011). *Fairness versus justice in language testing: The place of English literacy in the Australian Citizenship Test*. Language Assessment Quarterly, 8(2), 161-178.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103).

New York: American Council on Education and Macmillan.

Nagai, N., & O'Dwyer, F. (2011). *The actual and potential impacts of the CEFR on language education in Japan*. Synergies Europe, 6, 141-152.

Negishi, M., Tono, Y., & Fujita, Y. (2012). *A validation study of the CEFR levels of phrasal verbs in the English Vocabulary Profile*. English Pro□le Journal, 3, e3 doi:10.1017/S2041536212000037

Nematzadeh, A. (2018). Construct Irrelevant Factors and Test Validity: Investigating the Relationship among Gender, Age, Mother Tongue, Field of Study and TOEFL IBT ® Results. *Foreign Language Research Journal*, 8(1), 139-166. doi: 10.22059/jflr.2018.242996.405

Peachy, W. S. (2012). şen, Y.(2012). *The Common European Framework of Reference for Languages (CEFR) and English Language Teaching in Higher Education in Turkey*. In International Higher Education Symposium.

Piccardo, E. (2012). *Multidimensionality of assessment in the Common European Framework of References for Languages (CEFR)*. Les Cahiers de l'ILOB/OLBI Working Papers, 4, 37-54.

Karimi, F., Chalak, A., & Biria, R. (2019). *Pedagogical Utility of Pre-Listening Activities for Improving Iranian Elementary EFL Learners' Listening Comprehension*. *International Journal of Instruction*, 12(1), 1127-1140.

Kunnan, A. (2000). *Fairness and justice for all*. In A. J. Kunnan (Ed.), Fairness and validation in language assessment, Vol. 9, Studies in language testing (pp. 1–14). Cambridge, UK: UCLES/CUP.

Kunnan, A. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context* (pp. 27–48). Cambridge, England: UCLES/Cambridge University Press.

Kunnan, A. (2008). Towards a model of test evaluation: Using the test fairness and test context frameworks. In L. Taylor & C. Weir (Eds.), *Multilingualism and assessment* (pp. 229–51). Cambridge, England: UCLES/Cambridge University Press.

Le, H. T. T. (2018). Impacts of the CEFR-Aligned learning outcomes implementation on assessment practice. Hue University Journal of Science: Social Sciences and Humanities, 127(6B), 87-99.

Little, D. (2014). Learning, teaching, assessment: An exploration of their interdependence in the CEFR. In *5th international conference on teaching English as a foreign language assessment in ELT: Opportunities and challenges, FCSH, Lisbon new university, Portugal* (pp. 21-22).

Linacre, J. M. (2019). Winsteps[Computer program]. Chicago, IL: Winsteps.com.

Lundell, F., Lindqvist, C., & Edmonds, A. (2018). Productive Collocation Knowledge at Advanced CEFR Levels: Evidence from the Development of a Test for Advanced L2 French. Canadian Modern Language Review, 74(4), 627-649.

Martyniuk, W. (Ed.). (2010). *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual* (Vol. 33). Cambridge University Press.

Megyesi, B., Granstedt, L., Johansson, S., Prentice, J., Rosén, D., Schenström, C. J., ... &

Pérez de la Calle, S. (2018). The use of advertisement as didactic resource in the foreign language classroom according to sociocultural, linguistic, sociolinguistic and pragmatic aspects.

Razavipour, K. (2019). Philosophy of Research in Language Testing: Investigating Papers Published in Iranian, Peer-reviewed, Domestic Journals from 2008 to 2018. *Foreign Language Research Journal*, 9(3), 831-860. doi: 10.22059/jflr.2019.262709.535

Shakirova, D. S., Zamaletdinov, R. R., & Ashrapova, A. K. (2018). The Impact of the Cefr in Testing Tatar as a Foreign Language (A2 Level). The Journal of Social Sciences Research, 4, 36-39.

Shaw, S., & Imam, H. (2013). *Assessment of international students through the medium of English: Ensuring validity and fairness in content-based examinations*. Language Assessment Quarterly, 10(4), 452-475.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge university press.

Wu, J. R., & Wu, R. Y. (2007). Using the CEFR in Taiwan: The perspective of a local examination board. *The Language Training and Testing Center Annual Report*, *56*, 1-20.

Xi, X. (2008). Methods of test validation. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopediaof language and education. Vol. 7: Language testing and assessment* (2nd ed., pp. 177–96). New York, NY: Springer.

Xi, X. (2010). *How do we go about investigating test fairness?* Language Testing, 27, 147–170