Language Testing in Asia

RESEARCH Open Access

Assessing the content typicality and construct of Persian language proficiency test (PLPT) for non-Persian speakers: a corpus-informed study

Mahmood BijanKhan[1], Parvaneh ShayesteFar[2*] and Hassan Mohebbi[3,4]

*Correspondence:
parishayeste@cfu.ac.ir;
parishayeste@yahoo.com

[1] Department of Linguistics, Faculty of Literature and Humanities, University of Tehran, Tehran, Iran
[2] Department of English Language Teaching, Farhangian Teacher Education University, Tehran, Iran
[3] Tehran, Iran
[4] European Knowledge Development Institute, Ankara, Turkey

Abstract

Drawing on a growing body of research on the interface between corpus linguistics and second/foreign language testing and assessment, we adopted *Peykare*, a large-scale, annotated, Persian written language resource to evaluate the content (i.e., coverage and typicality) and construct validity of a Persian language proficiency test developed for certification of proficiency in Persian as a foreign language (PFL) of non-native speakers. Designed at the Research Center for Intelligent Signal Processing (RCISP), Peykare contains 35,058 text files over five linguistic varieties and 24 different registers of contemporary Persian. This study addresses how corpora, as rich database resources, can practically be applied to test validation purposes and insightfully inform the test life cycle. The results of content validity phase revealed evidence supporting content representativeness, relevance, and typicality of the test. The linkage between the corpus-extracted criterial features or parameters and those covered by the test was not, however, strongly evidenced by items measuring *ezafeh* constructions, homographs/homophones, PRO (proposition), and POST (postposition). The analysis of content typicality indicated chunks that did not closely conform to the corpus typical output. The construct validity phase, assessing the test hypothesized factor structure (i.e., hierarchical, unitary, correlated, and uncorrelated models) in two randomly split samples of PFL learners from Asian and European countries (N=121), showed that the correlated model fit the data best in both samples. The results supported the presence of distinctive factors of receptive skills, providing empirical evidence for score interpretations of the corpus-based test.

Keywords: Peykare, Corpus database resource, Persian as a foreign language (PFL), Content typicality, Corpus-informed tests, Persian language proficiency test (PLPT), Test validation

Introduction

Alongside the use of corpora data in second language teaching, there is a growing recognition in their use for second language testing and assessment (LTA). Since the 1990s, corpora have been increasingly used as a reference resource to identify the linguistic

features of native and learner usage suggesting aspects of language to test or to avoid (Hunston, 2022; Park, 2014). Over the time, corpus-based research, as an emerging area in LTA, has served as a basis to inform testing practices and decisions (Egbert, 2017; Römer, 2022). Given their potential contribution to designing and validating large-scale language tests such as second/foreign (L2) proficiency tests, for instance, corpora can inform not only what to assess as a language construct but also how to operationalize each construct or skill in specific tasks. Such information is essentially useful in describing what performance is typical at various proficiency (or can-do) levels; therefore, it will be potentially important for systematically comparing the linguistic features found in learner language with those associated with native users. Despite the growing popularity and availability of representative corpora as unique sources of authentic data for LTA (Taylor & Barker, 2008), to the best of our knowledge, few test makers have applied them in developing or validity evaluation of language proficiency tests. To narrow this gap, initiatives were taken by the present study.

The benefits of reference corpora are now established in LTA discipline, worldwide (see e.g., Cushing, 2017; Gyllstad & Snoder, 2021; Hughes, 2008; McCarthy, 2010). More recently, test developers have drawn on such corpora to develop and validate the way in which language proficiency construct is operationalized along specific levels or exemplified through level-specific linguistic demands. This mirrors how corpus evidence informs both the development of a list of level-specific linguistic indicators and their inclusion in a test to differentiate learners of different proficiency levels from each other or from the native speakers (Barker, 2010). Typical performance indicators associated with proficiency levels, also called Reference Level Descriptors (RLDs; Council of Europe, 2011), have been developed by the Association of Language Testers in Europe (ALTE, 2002) in terms of a set of "can-do" statements aligned to the Common European Framework of Reference (CEFR; Council of Europe, 2001). As an example, English Profile project (an interdisciplinary research program) uses the thirty-million-word Cambridge Learner Corpus to provide a set of RLDs for English for all six levels of CEFR, from A1 to C2.

Taking inspiration from corpus-attested patterns, CEFR provides a coherent framework for description of learners' proficiency ability (i.e., communicative competence) at each of the six levels (i.e., A1 to C2), for example, "Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension" (B2 level reading overall general ability; Council of Europe, 2001, p. 69). On the basis of such usage-based characteristics, CEFR, therefore, identifies linguistic features (i.e., RLDs) that are "criterial for distinguishing one level from the others" (Salamoura & Saville, 2009, p. 34). The key criterial features which are typically found across all CEFR levels are *lexical, grammatical, phonological, orthographic, semantic, sociolinguistic,* and *pragmatic* (see Saville & Hawkey, 2010). The framework primarily serves as a guide for self-assessment of language ability through calibrated scales (Council of Europe, 2001); nevertheless, when the insights taken into modeling proficiency tests are considered, it has appeared much influential in LTA over the last 15 years (see, e.g., North, 2014; Taylor & Jones, 2006). Although CEFR is not a "one-size-fits-all approach to measuring language ability" (Council of Europe, 2011, p. 9), it can be employed as a whole and adapted to the context of the test.

To design a CEFR-aligned proficiency test, test developers, therefore, need to initially establish the nature of proficiency at different levels (Barker, 2010; Hendriks, 2008) and make decisions about the illustrative features and descriptors that are criterial across these levels. Corpus data have both theoretical and practical contribution in identification, specification, and association of these typical linguistic exponents to each of the levels (Taylor & Barker, 2008). For instance, Kennedy and Thorp (2007) report how key features of IELTS written corpus including texts of different levels of writing performance led to the formulation of the band descriptors used to assess IELTS writing. Likewise, the British National Corpus (BNC), as another key reference corpus, was used by Cambridge ESOL to inform and validate the level-specific writing tasks (Saville, 2003). Large corpora like these, undoubtedly, can provide language testers with reliable information for each criterial feature, helping them through test designing process. Information about the ratio of the occurrence of a particular feature or RLD has been reported in the available literature as one of the key criteria to determine, select, and finally include specific criterial features or RLDs in a test (Beglar & Nation, 2013). In other words, *frequency measure criterion* forms a plausible direction (Alderson, 2007) and helps test designers explore key patterns that are common/less frequent at certain levels (Barker, 2010), thus guiding their inclusion in a test.

Despite suggested potential usefulness of corpora for test development and validation, their practical application has remained underexplored on an unprecedented scale (Taylor & Barker, 2008). To extend this line of research, Hunston (2022) argued for three interfaces between corpus linguistics and language testing, namely material selection and specification, development and validation, and scoring. In her view, corpora information and evidence can inform test validation, particularly in the aspects of test *content* and *construct* which, according to Hughes (2003), are important facets of validity. For instance, corpora extracted linguistic and discourse features help determine whether the language of test items reflect real-life language production (Weir & Milanovic, 2003), that is, whether the produced features are observed in native speakers' norms and usage in terms of *content*, *form,* and *frequency* of occurrence.

Although different methods were separately used to determine test validity before the 1980s, in the current conceptualization of validity, validation starts with collecting evidence, from different sources, in support of test interpretations and uses (see, e.g., Kane & Bridgeman, 2017; Messick, 1996). Mostly, test validity evidence can be collected from content relevance and content coverage as well as meaningfulness of test construct, viewed as "content" and "construct" validity, respectively. Arguing for the essentiality of construct validity in his unitary model of test validity, Messick (1993, 1996) considered *domain relevance* and *representativeness* not as the surface content of test tasks or items but the attribute, skill, or knowledge measured by the test tasks or items.

To address content relevance and content coverage of language proficiency tests, test specification, especially the range and nature of targeted linguistic features, should be aligned with evidence from non-test real-life language production. This can be ensured through introducing reference corpora into test designing and test validity where specified test items can be aligned with non-test linguistic points that are challenging to the learners in a contextualized test (Jamieson et al., 2000). When developing test contents, test designers ought to pay more attention to the patterns and frequency information of

these challenging items that are generally undermastered by language learners (Pan & Qian, 2017). Given that in a corpus-informed test specification, the most frequent words, patterns, and tendencies of language usage should receive importance and priority (Hunston, 2022), the concept of *content typicality*, as an essential facet of content validity, comes into the forefront. Content typicality, in Pan and Qian (2017) term, refers to the "frequently occurring linguistic features in the reference corpora" (p. 124).

Drawing upon such a corpus-driven test basis, in an earlier study, Bijankhan and ShayesteFar (2016) used a reference corpus of written Persian (Peykare, a 100-million-word corpus developed by Bijankhan et al., 2011) for the initial development of an academic version (AV) of a CEFR-aligned Persian language proficiency test (PLPT) and further examined the test listening construct. Nevertheless, the test has not been examined for its content validity neither for construct validity arguments of its models of receptive skills, in particular, the higher-order model of receptive skills. The PLPT-AV project was sponsored by the Ministry of Higher Education of Iran, locally called Ministry of Science, Research and Technology (MSRT), with the ultimate goal of developing a standard language proficiency test in Persian intended to assess Persian as a foreign language (PFL) ability of non-Persian applicants of Iranian universities. To provide a reliable and appropriate assessment of language performance of these learners, the corpus database was closely analyzed for its criterial features. Although tests for all four skills (reading, writing, listening, and speaking) were developed during the PLPT-AV project, the content and structure of the latent construct measured by its overall receptive skills were investigated in this study.

While endeavors were made, in some earlier studies, to develop Persian proficiency tests, for example, a Persian proficiency test designed by Ghonsooli (2010), there seems to be least evidence in support of the existence of a corpus-informed CEFR-based leveled test of Persian prior to 2016 when theoretical underpinnings and empirical evidence of the PLPT-AV listening section were provided by Bijankhan and ShayesteFar (2016). Situated within a broad corpus-informed LTA context, the present study, therefore, briefly represents attempts to address the recent initiatives in the development of the corpus-driven PLPT-AV, and then mainly examines both the content (i.e., content typicality and content relevance) and construct validity of its receptive modules.

This paper describes how Peykare, a large-scale database source in Persian language, can inform and provide evidence to test development and evaluation cycle. The remainder of this paper is organized as follows. "Corpus-informed language testing" section first presents a review of the available literature on corpus-informed LTA, and subsequently focuses on the main structure of our corpus, its different sources of evidence, and the architecture of our test along with the hypothesized factor structure of the test. "Method" section presents our methodology of applying the corpus data source for test evaluation followed by analytics results, discussions, and conclusion in "Results" and "Discussion" sections, respectively. Finally, "Conclusions" section presents some of the main implications and benefits that can be gained from this research.

Corpus-informed language testing

Over the past two decades, the development of corpus use in language testing has been evidenced by research findings, for example in studies conducted by Alderson (1996), Barker (2010), Callies and Götz (2015), Kennedy and Thorp (2007), Huang et al. (2018),

McCarthy (2010), Park (2014), and Taylor and Barker (2008). An increasing use of corpora in LTA occurred with systematic electronic collections of written and spoken data by institutions and examination boards. In the USA, for instance, the Michigan Corpus of Academic Spoken English (MICASE) was set up as an archive of American university speech (native and non-native) and used by the English Language Institute of University of Michigan (ELI-UM) for development and validation of examinations (see Simpson et al., 2002). Earlier in the 1990s, the EFL Division of the University of Cambridge Local Examinations Syndicate (UCLES EFL) and Cambridge University Press developed Cambridge Learner Corpus (CLC) as an archive of learning writing scripts, demographic, and score data. Having initially included three proficiency levels of general English tests (the Certificate of Proficiency in English, Certificate in Advanced English, and First Certificate in English-FCE), CLC expanded to include other domains and proficiency levels beyond these three English exams.

To date, both corpus linguistics and language testing scholars have discussed various theoretical implications and practical applications of the corpora in language testing. Aided by theoretical advances in the fields of corpus linguistics and LTA, corpus-informed testing can allow development of test materials, test schemes and standardizations, and comparative activities (Ball, 2001; Barker, 2004). Likewise, in the view of Alderson (1996) view, the use of corpora data can reveal much about test compilation and selection, preparation, and delivery of test results. As an example of corpora practical application, MICASE was used by the ELI-UM testing and certification division as a source of information for test writers to get informed about aspects of spoken English such as realistic speech rates occurring in the academic settings of the USA. In a similar vein, in the UK, a corpus-based checklist was developed to validate academic IELTS speaking tests in terms of communicative functions in different domains (see Brooks, 2001). Analyzing the original and corpus-based edited reading texts of FCE, in terms of their lexis and phraseologies, Hughes (2008) supported the view that corpus-informed test content is relevant to the non-test real-world situations.

Notwithstanding the initial use of corpora for describing linguistics aspects, such as individual words, word meanings, grammar, and collocations and phrases, corpus linguistics descriptions can be applied to various stages in a test lifecycle, from defining test purpose to designing and rating tests. In relation to test purpose, for instance, corpora can reveal much about learner ability levels showing what language learners can do at particular levels of proficiency. This dimension has a role in helping language testers to develop test materials and test rating scales and ensure what is tested at a certain level, whether overall or at item level, is relevant to what is required by language users for a particular situation such as educational, vocational, or professional (Barker, 2004, 2010). In this process, corpora can aid language test designers to more accurately identify and focus on criterial features or RLDs that seem to reveal how each proficiency level differs from adjacent levels (Salamoura & Saville, 2009). This will result into a valuable data source and a useful tool providing test designers with more precise linguistic and functional descriptors addressed by all CEFR levels and bringing to light more evidence about the nature of language proficiency at all these levels.

In test designing process, corpora are explored and analyzed to show which collocational patterning of the criterial features are frequent/typical or less common at

particular levels suggesting what learners can be expected to know at these levels. Thus, corpus-based data are the best way to provide reliable frequency measures (Alderson, 2007) guiding test specifications and the inclusion of criterial features in tests. Such data, additionally, indicate the most frequent errors (in terms of words or collocational pairings) which can be used as appropriate items, i.e., answers or distractors in tests. As an example, the new listening items of the Examination for the Certificate of Proficiency in English, a high-level test, were developed based on the MICASE word frequencies (Barker, 2010). Purpura (2004) also highlights the frequency information and distribution patterns of lexical items as important sources for determining, examining, and assessing the contents of a test. Consistent with this view, Pan and Qian (2017) provided evidentiary basis for tackling content validity of the National Matriculation English Test (NMET) through corpora-incorporated data. Besides, according to Hawkey and Barker (2004) perspective, frequency of occurrence obtained through corpus analysis mirrors how to triangulate corpus methodologies to inform test life cycle and validation procedures. Given that the CEFR descriptors and statements are often too global and underspecified, such a corpus-driven frequency information can add "grammatical and lexical details to CEFR's functional characterizations of different levels" (Hawkins & Filipovic, 2012, p. 5) and quantify the criterial RLDs needed to distinguish between these ability levels.

Notwithstanding the available evidence suggesting that corpus linguistics and corpus-driven approaches have practical implications for LTA, the literature is not yet fully developed on high-stakes tests of L2 proficiency intended to measure non-native speakers' communicative language ability. To narrow this gap, the present study therefore took insights from *Peykare*, a contemporary Persian corpus, into operationalizing and validating the content and constructs of the PLPT-AV receptive modules (i.e., reading and listening sections) intended to measure PFL ability of non-Persian speakers. To our knowledge, there are no reports on a Persian proficiency test centered on corpus-informed method for both linking to CEFR levels (A2-C2 in this study) and validating the test content and structure.

Peykare: a corpus of contemporary Persian (2006 onwards)

Built as "a written language resource for the contemporary Persian" (Bijankhan et al., 2011, p. 143), *Peykare* is a large core corpus (100 million words, 35,058 texts) designed at the Research Center for Intelligent Signal Processing (RCISP). The corpus is representative of five linguistic varieties (Standard-informal, Standard-formal, Super-Standard-informal, Super-Standard-formal, and Sub-Standard-informal) and 24 registers that Persian speakers use and encounter in major disciplines (e.g., natural sciences, humanities, and arts). Collected from naturally occurring discourse of different academic, institutional, or constitutional registers (e.g., education, regulations, manuals), the corpus texts include "texts written to be read" (87%) and "texts written to be spoken/listened" (13%). Annotated Peykare has been closely searched through *Searchdata* tool to look for its syntactic and morphological resources (Fig. 1 shows an example of annotated trigger for word "پرسشنامه"/questionnaire (128 hits in 10 million words)). The outcome, according to Bijankhan et al. (2011), was the emergence of more than a dozen of general parameters, including *conditional*, *relative*, *complement*, and *passive structures; articles,*

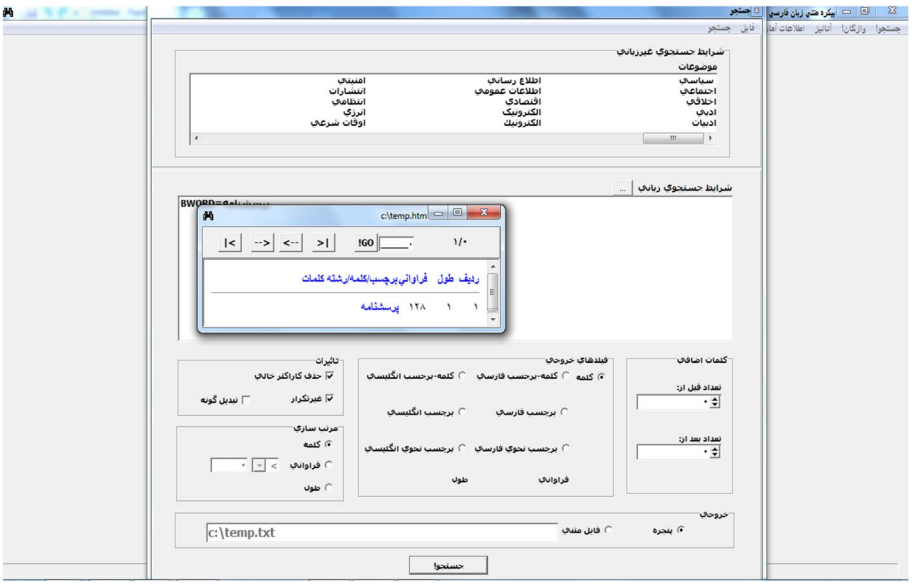

Fig. 1 An example of annotated trigger by Peykare's Searchdata

question words constructions, noun, *verb*, *adjective*, *adverb*, *preposition,* and *pronoun constructions;* and *homographs/homophones,* all in forms of monograms or strings of words of collocated bigrams, trigrams, or in general, *n*-grams.

To obtain the frequencies of occurrence of each parameter, all words with their tags were sorted in descending order. Accordingly, Peykare's parameters were selected and included within the PLP-AV tasks and item contents, with the least frequencies for C1-C2 levels, the moderate for B1-B2, and the highest frequencies for A levels. Of note, an additional parameter was taken as the basis for further text analysis due to its typicality in Peykare: "the parameter of *Ezafe".* In the Indo-European languages like Persian and Pashto, Ezafe is used as a linking element, an enclitic pronounced /*e*/, to link the head of a phrase to its modifiers and disambiguate the boundary of a syntactic phrase (Karimi, 2007). For example, the phrase "*divār-**e**Ān kelass-**e**Bozorg*" (in English: the wall of that big class) has two Ezafe constructions in text processing namely [N *EZ*][1] and [*DET* N *EZ* AJ][2]. A frequency counting of parts-of-speech categories showed that almost 20% of words in contemporary Persian texts "include words with Ezafe while no orthographic symbol is used to refer to it" (Bijankhan et al., 2011, p. 157). The highly occurring Ezafe constructions in Persian are Noun+Adjective (23.58%) and Noun+Noun (22.24%). The moderate occurrences were found for three-word constructions such as Noun+Noun+Noun (8.84%), Noun+Noun+Adjective (6.4%), or Noun+Adjective+Noun (4.20%) while the least frequencies of occurrences were observed for Noun+Determiner (2.60%), Pronoun+Noun (2.43%), or four-gram constructions such as Noun+Noun+Noun+Noun (2.17%) and Noun+Noun+Noun+Adjective (1.44%). Since Persian has no overt orthographic Ezafe

[1] The phrase "*divār-**e**Ān kelass-**e**Bozorg*" has two Ezafe constructions in text processing: [N EZ]: [NP[N[*N divār*]][EZ *e*]; and

[2] [DET N EZ AJ]]: [NP[DET *Aan*][N[N kelass]][EZ *e*][AJ *Bozorg*]]]. "One Ezafe construction is theoretically embedded within another [N EZ [DET N EZ AJ]]" (Bijankhan, Sheykhzadegan, Bahrani, & Ghayoomi, 2011, p. 157).

symbol, correctly recognizing this typical construction in longer word sequences will become hard. The same frequency criterion applied to the other parameters was followed for the selection and inclusion of Ezafe parameter in all test tasks and items.

From an applied perspective, compared to the existing examination data sources such as available PFL books or electronic sources, Peykare therefore offers practical benefits such as a broad coverage of lexicon-grammar parameters typically found in authentic forms of Persian language use.

PLPT-AV structure

Following a 2-year study phase of the test, its development process was set up with triangulated sources of data from (a) Persian language education policy documents, (b) Peykare corpus, (c) available materials on Teaching Persian to the Speakers of Other Languages (TPSOL) or what is locally known in Iran as *AZFA* (*Amooszesh-e Zaban-e FARSI*), (d) TPSOL/AZFA instructors and language test experts' theoretical knowledge and practical experience, and (e) international language proficiency tests like TOEFL, IELTS, and JLPT (Japanese Language Proficiency Test). The convergence of these sources led to the skill (reading, listening, writing, and speaking) and task specifications. Since a full and detailed report on multiple stages adopted for the construction of the test cannot be included within the present limited space, only major procedures are briefly described here.

The construction, standard setting, and modes of the tests were determined by the PLPT-AV development committee. Three linguists, two professional TPSOL experts from AZFA Institute of Tehran university, also known as *Dehkhoda Lexicon Institute and International Center for Persian Studies* (henceforth called Dehkhoda Lexicon Institute), one computational linguist and three postgraduate students, and the present researchers (the current dean of Dehkhoda Lexicon Institute, who is a full professor in General linguistics and AZFA, and an LTA expert with years of L2 teaching) were actively involved in the project. Content coverage and content representativeness of the test tasks or items (i.e., content standards), performance standards in terms of the CEFR scales, numbers and types of the tasks, relative weights of the skills and tasks, and item specifications (e.g., detailing about item acceptable vocabulary, syntax and content limits, item numbers, initial item reviewing) together with the scoring/rating procedures were all documented in the test specification phase.

The point of departure for test specification process was Bachman (1990) Communicative Language Ability (CLA) model taken as a nested model of *linguistic* (i.e., lexical, grammatical, semantic, phonological, orthographic), *sociolinguistic* (i.e., register differences, politeness conventions, and linguistic markers of social relations), and *pragmatic* (discourse and functional) *competences*. When aligned with the CEFR, each competence was described with a set of relevant criterial features/RLDs or can-do statements across six levels. For instance, an independent Persian language learner "Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension" (B1 level Reading overall general ability, Council of Europe, 2001, p. 69; Domain: *education*; function: *recommendation letter from a Persian Academic Staff in Iran*).

In the second phase, the methodology of standard setting was adopted by the committee experts to define the "levels" of proficiency in Persian and the corresponding cut scores. These judges (a) implemented a shared understanding of the CEFR levels, (b) worked within the exemplar performance and test tasks to achieve more adequate understanding of the levels, and (c) maintained the standards in a consistent fashion. The results of the standard setting procedures came into a total scale of 0 to 120 points, with each section (reading, listening, speaking, and writing) receiving a scaled score from 0 to 30.

Checking on the clarity and comprehensibility of the test items and rubrics as well as estimating the required time were followed in a phase after. The outcome was the elimination of problematic items such as perceived ambiguous items (three items), difficult texts (e.g., one literary text that needed specialized knowledge to comprehend), or complex structures with multi-unit Ezafe constructions that could not fit the level ability indicators.

In pretesting the test to a sample of foreign students with different L1s and nationalities, information on psychometric characteristics were obtained. Cronbach's alpha consistency was estimated and the reliability coefficient was found to be .82.

PLPT-AV receptive skills construct: the hypothesized factor structure models

The score reporting format of the PLPT-AV was used to hypothesize the receptive skill construct in form of its underlying factor models. PLPT-AV score report shows a single overall score along with a score for each of the reading and listening skills. Following this view that "the use of a single total score assumes that a single *higher-order* or *hierarchical factor* underlies performance on both the listening and reading sections of the test, whereas the use of separate scores assumes that distinctive factors of listening and reading skills are involved" (In'nami & Koizumi, 2011, p. 134), separate scores were used to assume the presence of distinctive factors of reading and listening skills in the PLPT-AV. On such a base, a higher-order/hierarchical factor structure was proposed to underlie the PLPT-AV performance on the reading and listening skills, where both skills are involved in and influenced by the total receptive skill. Such a hypothesized structure concurs with the literature on language ability as a hierarchical model (see Bachman & Palmer, 1982).

Yet, in a study on the University of California Los Angeles English as a Second Language Placement Examination scores, Oller Jr. (1983) reported a single trait model of language ability. His unitary trait model was later disproved by those studies that analyzed students' test performance through rigorous methods such as confirmatory factor analysis reporting L2 language ability as a correlated trait model (see In'nami & Koizumi, 2011) or higher-order ability model. Moreover, a close relationship between reading and listening processes has also been supported by psycholinguistic studies reporting a common process underlying the two separable skills (e.g., Hirai, 1999). Nevertheless, in another study, Wilson (2000) found that listening ability is not correlated with other language abilities such as speaking or reading. Aligned with these inconsistent views, we hypothesized that reading and listening skills of the PLPT-AV are either (a) hierarchically structured, (b) separable and uncorrelated, (c) separable but correlated, or (d) inseparable (unitary model).

In addition to investigating the aspects of content validity of the test receptive skills, this study, therefore, aimed to investigate whether the higher-order or componential model assumed in the PLPT-AV fit the test performance data better than the unitary, correlated, and uncorrelated models. With these aims in mind, we addressed the following research questions:

(1) To what extent do the PLPT-AV linguistic parameters represent the target parameters in the Peykare data source? (*Content validity: coverage and typicality)*
(2) Does the higher-order model assumed in the Peykare-driven PLPT-AV fit the data better than the correlated, uncorrelated, and unitary models? *(Construct validity)*

Method

Participants

The database used for evaluation of the utility and efficiency of Peykare as a natural user-generated content for LTA represents performance scores of 121 Persian language learners from 16 European and Asian countries, with the average age of 28.3, ranging from 19 to 51. The sample consisted of 82 (68%) males and 39 (32%) females who had enrolled in either a graduate or undergraduate program in University of Tehran or Dehkhoda Lexicon Institute during the study. The largest groups (76%) were from Iran's neighboring countries (i.e., Asian countries) and the rest of participants (24%) were from European countries. It was almost representative of the Persian language learners' population, where the largest groups are from the Asian countries.

The test performance data were derived from the newly designed PLPT-AV tests administered among the participants, in consultation with authorities of University of Tehran and Dehkhoda Lexicon Institute. Care was taken in the administration of the tests: test materials and booklets were kept secure, audio systems were checked along with the physical setting, and proctors received short training. This included information about administration timetable, administration guidelines (whether to admit latecomers, how to behave during the test, guiding test-takers through their seats, etc.), delivering the test booklets to the test-takers, and returning them back to the PLPT-AV committee.

The expert panel for content alignment judgment included Peykare's complier and designer who had years of expertise in AZFA and computational linguistics, and a university professor specialized in LTA and L2 teaching.

Instrument and procedures

The outcome of the PLPT-AV project was a 120-point test measuring all four PFL skills through four 30-point measures (listening, reading, writing and speaking sections). As to the receptive skills, the listening skill was defined as an integrated-skill domain including listening-reading and reading tasks; however, the decision was made based on the modality of output. The listening booklet included six communicative events (e.g., asking for services, radio interviews, lectures) of specific length and difficulty levels, and in forms of dialogue, conversation, and monologue tasks. Overall, the 30-item listening section measured the ability to (a) locate straightforward factual information; (b)

infer gist and purpose of short spoken speech on explicit information; (c) infer gist and purpose of extended spoken speech on explicit information; (d) understand details in extended spoken texts; (e) understand and follow extended speech on abstract and complex topics; and (f) understand all forms of spoken language (i.e., live or broadcast texts of implicit explicit meaning) delivered at fast speed. The tasks, paced by a voice-recorder and compact disk, lasted for 50 min.

The reading skill booklet was another 30-item measure consisting of six passages of five items for measuring the ability to (a) understand different texts of varying difficulty level and lengths (ranging from simple notices, timetables, personal letters to articles, technical instructions, literary, and complex texts); (b) understand details in lengthy and complex texts; (c) locate and use specific information and reference sources selectively; (d) make inferences in written texts; (e) understand straightforward factual texts; and (f) understand gist and purpose of short texts.

Regarding the length and difficulty, the tasks were not equivalent; they were designed and sequenced in an increasing length and difficulty level requiring both linguistic and non-linguistic knowledge to work interactively to produce comprehension. All items of the two receptive skills were objectively marked. Cronbach's alpha reliability coefficients were 0.87 and 0.79 for the reading and listening tests, with the reported means of 21.43 and 18.70, respectively.

As to the content validation phase of the study, the corpus of contemporary modern Persian (i.e., Peykare as discussed in 2.1 above) was used. Peykare is fairly distributed in terms of register coverage (24 registers encountered in disciplines of arts, humanities, and natural sciences) as well as genre coverage. The corpus includes written texts; however, spoken corpora have also been produced in projects such as FARSDAT, TFARSDAT, and the Large FARSDAT in the country. In order to make comparisons between the test and Peykare's content, Searchdata tool was used looking for the most frequent morphological, syntactic, and collocated structures via sorting parameters with their tags in descending order. As to the construct validation phase of the study, a model comparison approach using multiple structural equation modeling (SEMs) was employed to examine how well each model assumed in the PLPT-AV receptive skills designed based on Peykare's linguistic features fit the non-Persian speakers' test performance data.

Analysis

To address research question 1, an alignment judgment approach was taken by the expert panel to align the assessed parameters with those targeted by the corpus. Since this question concerned typicality as well, further statistical analyses of corpus data were performed to check the correct answers and distractors of each test item against frequency information obtained from the search engines of the corpus. We judged the item might not be typical if its content frequency was lower than those in the corpus or those of distractors.

Regarding the test takers' performance data, both the scores and the percentage of responses in each subskill were available for the analysis. To address the second research question, the test items were used as measures of receptive skill construct, thus, their scores were used for twelve observed variables in each model. The factor models hypothesized on the basis of available research and theory and the PLPT-AV scoring process

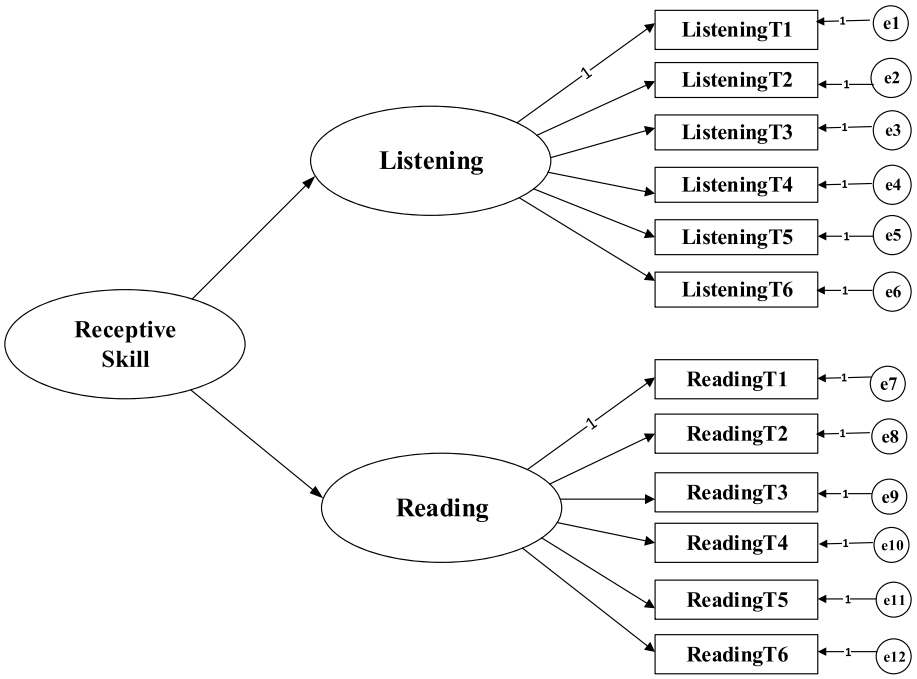

Fig. 2 Hierarchical/higher-order model

were as follows: a hierarchical trait model (Fig. 2), an uncorrelated trait model (uncorrelated reading and listening skills), a correlated trait model (Fig. 3), and a unitary trait model (with no distinct reading and listening subskills).

In order to test multiple structural models against each other, the PLPT-AV data were randomly split into two halves, following Byrne (2010) and MacCallum et al. (1994). Then, confirmatory factor analyses with AMOS (version 18: Arbuckle, 2009) were employed to examine the factor model of the test in each sample (half). Model parameters were estimated through Maximum likelihood method. The normed chi-square (shown by CMIN/DF), a non-significant chi-square (χ^2) and other goodness-of-fit indices (GFIs) that test the consistency of the proposed models with the pattern of covariations among the observed variables, and the Root Mean Square Error of Approximation (RMSEA), as another informative index of how close each model corresponds with the data, were inspected. Kurtosis and Skewness values were checked for the normality of distribution of the variables.

Results

Content validation results

The expert panel first closely analyzed Peykare to find the most frequently occurring parameters and subsequently aligned the targeted parameters in the corpus with those parameters and features assessed by the items to address the issue of test content. The analysis resulted into identification of several lexical, phonological/graphological, and morphosysntactic parameters as the most typical ones, such as (a) parts of speech (e.g., noun, verb, preposition and postposition), (b) Ezafe construction, (c) homograph/homophone (e.g., "مردم"/*mardom*/ (people) or /*mord+am*/ "I died"), (d)

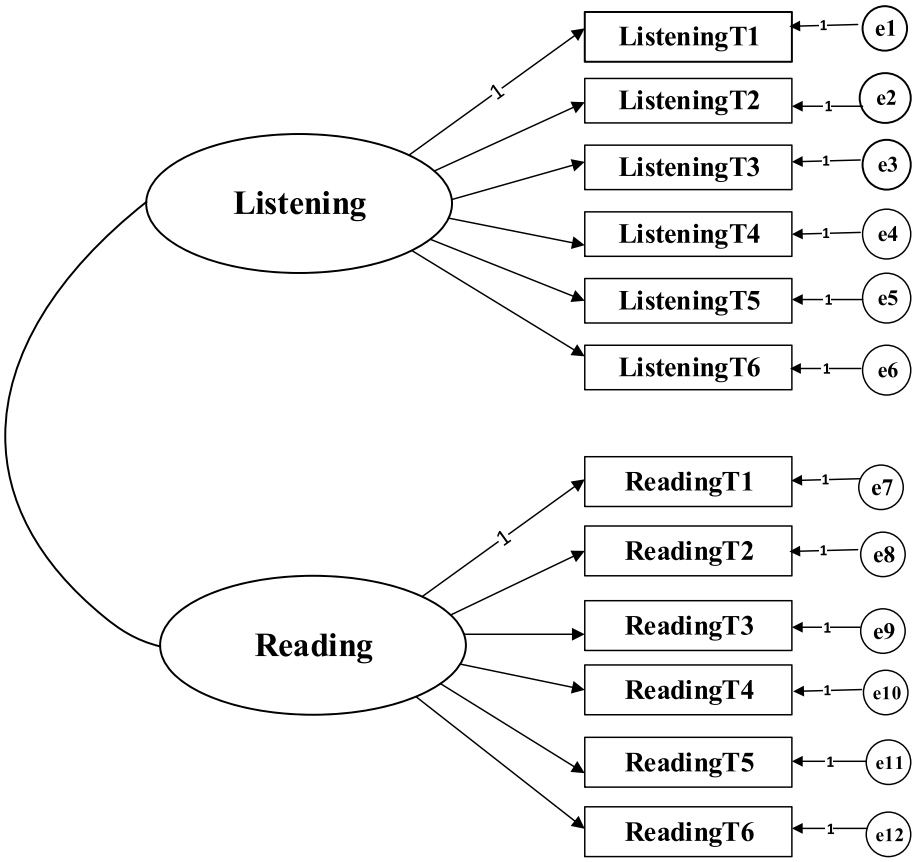

Fig. 3 Correlated model. Note: The ovals represent latent variables/factors; the circles, measurement errors/residuals; and the rectangles, observed variables

adverbial, (e) passive, and (f) conditional clauses. (See Table 1 for the percentages of the target parameters in the corpus and test specifications). Regarding POS, Table 1 shows the highly frequent occurrences for the categories of nouns (N: 39%), prepositions (P: 11.25%), punctuation (PUNC: 10.27%), adjective (AJ: 9.27%), verb (V: 8.89%), and conjunction (CONJ: 8.48%). The expert panel matched the POS and Ezafe construction items with their corresponding categories in Peykare and confirmed all these items were realized in the test. In their view, what was measured through these items was in consistency with what had been intended in the test specifications. Drawing upon Table 1, it was found that the test items were related to and covered the specified lexico-grammatical parameters except PRO and POST in Listening measure. A clear example of POST is "را" /*ra*¯/ which fuses with some major parameters to show definite marker after the nouns (/ 'گفتم را وی'*I said to her*/).

The lower frequency rate of passive clauses and homographs/homophones (in Reading and Listening tests) is notable. Compared to the reported proportion in the corpus (60%), the total items observed under passive clause parameter account for 42% of all the test items. Likewise, 49.5% (Reading) and 33.3% (Listening) of all the test items were accounted for by homographs/homophones parameter, as compared to 58.5% rate of occurrence in the corpus. Moreover, regarding the Listening test, the

Table 1 Distribution of parameters in each targeted parameter of Persian knowledge and the test

Parameters	Frequency and percentage		
	Reading test	Listening test	Peykare (%)
POS			
Noun (N)	12 (40%)	13 (43%)	39.74
Preposition (P)	4 (13%)	6 (20.3%)	11.25
Adjective (AJ)	3 (10%)	3 (10%)	9.27
Verb (V)	3 (10%)	2 (6.6%)	8.89
Conjunction (CONJ)	2 (10%)	2 (6.6%)	8.48
Number (NUM)	1 (3.3%)	2 (6.6%)	3.13
Pronoun (PRO)	2 (6.6%)	0 (0%)	2.58
Determiner (Det)	1 (3.3%)	1 (3.3%)	2.50
Adverb (ADV)	1 (3.3%)	1 (3.3%)	1.84
Postposition (POST)	1 (3.3%)	0 (0%)	1.47
Homograph/Homophone			
Non-lexical (e.g., verbal 3rd person, preterite 3rd person, preterite/perfect 1st person, possessive 2nd person,)	15 (49.5 %)	10 (33.3 %)	58.50
Ezafe construction			
N+N	13 (43.3%)	11 (36.6%)	33.24
N+AJ	7 (23.3%)	3 (10%)	23.58
N+N+N	5 (16.5%)	8 (26.4%)	8.84
AJ+N	1 (3.3%)	2 (6.6%)	5.38
N+N+N+N	2 (6.6%)	4 (13.3%)	2.17
N+N+N+AJ	2 (6.6%)	2 (6.6%)	1.44
Clauses (co-ordinate, subordinate, predictive, …)			
Conditional	2 (6.6%)	2 (6.6%)	7.7
Passive	14 (42%)	12 (40%)	60
Adverbial	1 (3.3%)	3 (10%)	2

POS parts of speech

proportion of P, and NAJ, NNN, and NNNN in Ezafe constructions did not closely cover the specified proportion in the corpus. Though the parameters, on average, were covered in the PLPT-AV, the former appeared to be underassessed whereas the later were overassessed. Regarding the content coverage, the expert panel found certain degrees of correspondence between the observed parameters and the already specified ones. However, the extent of divergence observed is worth further consideration.

In order to further probe the test content validity, content typicality "as a newly proposed facet of content validity" (Pan & Qian, 2017, p. 130) was used. In so doing, the degree to which the assessed parameters and their subcategories were typical compared with their presence in real-life language use was examined. For this purpose, correct responses and distractors, together with their adjacent words, were concordanced in Peykare. The results revealed degrees of inconsistency along two lines. First, typicality was observed in relation to the contents of the test levels. We found cases where the correct answers included chunks that were as typical as expected yet not typical at certain proficiency levels, particularly A1 and A2 levels. For instance, "applicant" and "free from" each can be stated by two forms in Persian, /motæqa:zi:/ or /dɑ:vtalæb/ and /fa:req æz/ or /a:za:d æz/ respectively. The latter form in each pair is less frequent than the former, therefore, not typical or appropriate for low-level language learners such as A-level testees. Likewise, chunks such

as /maqu:la:te ha:eze æhami:jæt/ and /mozu:?a:te muhem/ "important issues" (in Listening test) hit across Peykare, yet with different frequencies which make one form more appropriate for intermediate or advanced level testees than for the beginners.

Further flaws of content typicality were found with chunks involving collocational prepositions, for instance [N P N N] strings. As illustrated by Example 1, the chunk /pa: dar ærsejeh vudʒu:d guza:shtæn / "to come into the existence" is both grammatically and semantically acceptable, yet when the preposition is concerned, it is not as typical as presumed. The reference corpus does provide evidence in support of this concordance /pa: be ærsejeh vudʒu:d guza:ʃte ænd/ (choice b). Yet frequency profiling of choices c (ج) and d (د) yielded lower degree of typicality (45 hits) showing inefficiency of these options or distractors. It was found that the most frequently occurring chunk would be ""پا به عرصه وجود گذاشتن" 825) hits). Expert panels found 5 similar examples like this in PLPT-AV Reading section.

Example 1(# Reading skill)

'امروزه به‌دنبال رویکرد جهانی شدن، چالش‌های متفاوتی پا ..... گذاشته‌اند'.
الف) در عرصه وجود ب) به عرصه وجود ج) در حیاط هستی د) به حیاط هستی

[emru:zeh be dunba:le ru:jkærde jaha:ni: ʃ udæn tʃa:leʃha:je mutafa:veti: pa: ..... guza:ʃte ænd]

a) dar ærsejeh vudʒu:d *b) bə ærsejeh vudʒu:d c) dar haja:te hæsti: d) bə haja:te hæsti:

English: /Today, as a consequence of globalization trend, various challenges have come into the ..... .

Conversely, there were other cases in which the frequency profiling yielded higher degree of typicality for distractors than the correct answer. As Example 2 illustrates, the frequency of "یک سوم"/ one-third/ (122 hits) and "وقت...را نگیرد" /not taking time/ chunks (107 hits) is much higher than that of "نقاط مثبت را برجسته نماید" /highlighting the positive points/ as the correct answer (23 hits). Even though this chunk is grammatically and semantically acceptable, it is not naturally occurring as frequently as the distractors. Yet, since the correct answers to receptive skill tests are text-bound, they are inevitably retained. However, preferences for retaining a less frequently used chunk might cause complication. In this particular case, although the distractors are grammatically appropriate and do not fail to provide evidence for naturally used expressions, their efficiency or distracting power seems to be limited. Given this, test developers should replace distractors with either V-constructions such as choice c (ج) or d (د) which fit the stem or with expressions with similar frequencies as that of the correct answer.

Example 2 (# Reading skill)

با توجه به متن، طول مناسب برای یک توصیه نامه .....است.
الف) یک سوم یک صفحه کامل
ب) بیشتر از یک صفحه کامل
ج) به اندازه‌ای که نقاط مثبت را برجسته کند
د) کوتاه، به اندازه‌ای که وقت خواننده را نگیرد

[ba tavajoh be matn, tu:l-e muna:seb bæra:je jek tɔ:sijeh na:meh..... æst]

(a) jek sevoum-e jek sæfheje-e kæmel
(b) bi:ʃ tæræz jek sæfhejeh ka:mel
(c) be ænda:ze-ei: ke nuqa:a:t-e musbæt ra: bærʤeste kunæd
(d) ku:ta:h, be ænda:ze-ei: ke væqte xa:nænde ra: nægi:ræd

English: /According to the passage, the optimal length for a recommendation letter .....

Construct validation results

In response to research question 2, the study examined the degree to which certain constructs account for the PLPT-AV performance. Because empirical links are required to support the relationships between the intended score interpretation and the measured constructs (In'nami & Koizumi, 2011; Messick, 1996), a closer examination of the performance score interpretation as represented by the PLPT-AV would help understand how the results relate to the test constructs.

Descriptive statistics

As displayed in Table 2, all kurtosis and skewness values are within |3.30| (z score at p<.01), suggesting no violation to the univariate normality assumptions of the data. Mardia's coefficient was also checked for multivariate normality and the obtained value was below the recommended value of 20.00 (Harrington, 2009), indicating multivariate normality of the data.

SEM analysis: testing the four hypothesized models

The study tested the extent to which the four hypothesized models of the test components (constructs) are consistent with the obtained data. The following widely used criteria were used to assess the model fit: normed chi-square (CMIN/DF, χ^2/df) >1 & <3; Comparative Fit Index (CFI) and goodness-of-fit index (GFI) ≥ .90; p of individual observed variables <.05; RMSEA ≤.08; and χ^2 test of close fit (Byrne, 2006). For the χ^2 test of close fit (i.e., p-close), the hypothesis states that RMSEA is > 0.05, if the value is > 0.05, then it can be concluded that the model fit is close (Kline, 2011).

The results in Table 3 indicate that the chi-square statistic (χ^2=86.70, df=52, p<.05 for sample 1; χ^2=76.07, df=52, p<.05 for sample 2), CMIN/DF (1.6 for sample 1; 1.4 for sample 2), RMSEA (.08 in both samples), GFI (.86 for sample 1; .91 for sample 2), and p-close (.89 for sample 2; .022 for sample 1) values of the higher-order model are more acceptable than those of the unitary and uncorrelated models in both samples. Nevertheless, Table 3 shows that the goodness-of-fit indices of the correlated model were better than those of the higher-order, uncorrelated and unitary models, so the correlated model was the best model for the present data, showing an interpretable and meaningful model for Persian language skills of listening and reading. The model factor loadings were statistically significant (p<.05), ranging from .27 to .84 and .26 to .83 for sample 1 and sample 2, respectively. Although only one comparison statistic of unitary model, i.e., AIC (sample 1) was found more acceptable than that of the other models, when other goodness-of-fit values are considered, this model does not seem appropriate.

Table 2 Descriptive statistics of each sample

	Minimum	Maximum	Mean	Std. deviation	Kurtosis	Skewness
Sample 1						
Listening T1	1	5	4.43	.908	1.991	−1.687
Listening T2	1	5	3.56	1.293	−.842	−.444
Listening T3	1	5	3.51	.999	.390	−.889
Listening T4	0	5	2.88	1.316	−.498	−.286
Listening T5	1	5	3.35	1.435	−1.306	−.292
Listening T6	0	5	2.26	1.764	−1.634	−.307
Reading T1	1	5	4.21	1.151	1.899	−1.614
Reading T2	1	5	3.46	1.395	−.208	−.780
Reading T3	1	5	3.15	1.560	−1.064	−.341
Reading T4	0	5	2.81	1.730	−1.140	−.394
Reading T5	1	5	2.40	1.879	−1.657	−.037
Reading T6	1	5	2.10	1.868	−1.216	.45
Sample 2						
Listening T1	1	5	4.43	.785	1.462	−1.379
Listening T2	0	5	3.78	1.051	1.398	−.965
Listening T3	1	5	3.98	.812	1.980	−.948
Listening T4	1	5	3.18	1.467	−.845	−.362
Listening T5	1	5	3.40	1.278	−.972	−.248
Listening T6	1	5	3.26	.954	−.644	−.083
Reading T1	1	5	4.30	.808	1.961	−1.404
Reading T2	1	5	3.73	1.205	−.855	−.543
Reading T3	1	5	2.88	1.249	−.907	.066
Reading T4	1	5	2.83	1.209	−.636	.153
Reading T5	0	5	3.19	1.20	.260	−.655
Reading T6	0	5	2.01	1.518	−.894	.382

The minimum and maximum scores for each subskill are zero to 5 (these are raw scores, not in form of percentage)

Table 3 Goodness-of-fit indices of the four hypothesized models

Model	Fit statistics								
	χ^2	*df*	CMIN/ DF (1>, 3<)	CFI (≥.90)	GFI (≥.90)	RMSEA (≤.08)	AIC (the lower)	BIC (the lower)	*p*-close
Sample 1									
Unitary	95.68	53	1.7	.88	.79	.15	143.68	193.94	.006
Higher-order	86.70	52	1.6	.89	.89	.08	136.70	189.06	.022
Uncorrelated	94.53	53	1.7	.77	.81	.11	142.53	192.80	.007
Correlated	45.61	52	1.5	.91	.91	.07	124.88	168.86	.70
Sample 2									
Unitary	78.62	53	1.4	.86	.86	.05	108.62	158.88	.52
Higher-order	76.07	52	1.4	.89	.89	.08	124.07	174.33	.90
Uncorrelated	86.62	53	1.6	.79	.82	.09	126.42	168.31	.89
Correlated	65.61	52	1.2	.90	.91	.05	111.61	158.90	.71

Df/DF degrees of freedom, χ^2/*CMIN* chi-square, *CMIN/DF* χ^2/df

Table 4 Chi-square difference test results (correlated model Vs. unitary and uncorrelated models)

	χ^2 difference	df difference	Significance
Sample 1			
Vs. Uncorrelated	48.92	1	.01
Vs. Unitary	50.07	1	.01
Sample 2			
Vs. Uncorrelated	21.01	1	.01
Vs. Unitary	13.01	1	.01

df degrees of freedom

Overall, the results show that the unitary, uncorrelated, and higher-order models were less favorable than the correlated model. The uncorrelated model yielded poor fit across the samples. Furthermore, the existing literature provides less evidence in support of the unitary model (Sang et al., 1986). The analysis of the items loading on the two latent factors, factor loadings, and *p*-values of the correlated model could well support the distinctive constructs of reading and listening, i.e., the correlated model with two separate factors of reading and listening was confirmed. The path coefficients of the observed variables (ReadingT1 to ReadingT2, ListeningT1 to ListeningT6) to the corresponding factors of reading and listening were moderate to high (.31 to .84) and the correlation between the two factors was acceptable (.70), though less than .90; therefore, the two factors of reading and listening can be considered significantly distinct from each other. A further comparison was made between the fit of the correlated model and that of the unitary and uncorrelated models using *chi-square difference* tests. As shown in Table 4, the fit indices for the unitary and uncorrelated models were poor across both samples.

Overall, the results for Model 3 showed satisfactory model-data fit and support the acceptability of the model. The magnitudes of the factor loadings shown in Fig. 4 indicate that the tasks are reasonably good indicators of their associated factors. As to the factor loadings of different magnitudes, the largest loadings on the reading factor were for ReadingT6 (.84) and ReadingT5 (.61), and the largest loadings on the listening factor were for ListeningT2 (.73) and ListeningT6 (.71). Also, the magnitudes of their error variance are not substantial (.26 and.48 for ListeningT6 and ListeningT2, and .27 and .66 for ReadingT5 and ReadingT6). This suggests that the four tasks relatively serve as better measures of the two ability constructs. A further look at the loading magnitudes of other tasks indicated no loadings below .35, suggesting that the tasks items responsible for different parameters carry justifiable weightings and measure their constructs. A synthesis of the results obtained through estimation of the factor loadings, error variances, and the substantial correlation between the reading and listening factors (although less than .90) provides evidence for the presence of factorially distinct but correlated model. Taken as a whole, the results revealed that the receptive skills, as distinct constructs, account for performance on the test.

Discussion

In the last two decades, corpus linguistics has been increasingly used to inform LTA research and practice (Egbert, 2017). One major benefit of using corpora in LTA lies in their capacity for detailed analysis of linguistic features that distinguish language use across contexts, language users, genres, and proficiency levels (Cushing, 2017). Such

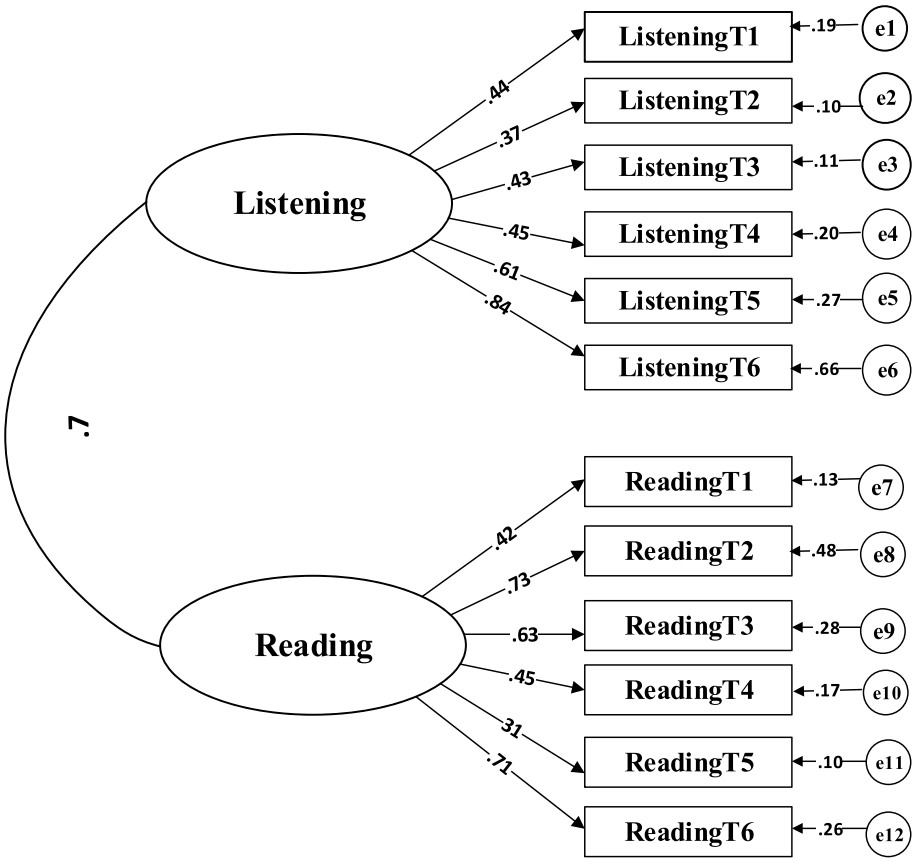

Fig. 4 Final model of correlated receptive skills. Note: All factor loadings were significant. The parameter estimates are standardized

information is particularly useful for identification and description of the criterial features and parameters that distinguish one proficiency level from the others. Given this, corpus linguistics has implications for test development decisions, such as theoretical and operational definition of test constructs, specifications, and selection of test content, for example, spoken or written texts for listening and reading tests, and test validation (see, e.g., Alderson, 1996; Barker, 2010, 2014; Batsuren et al., 2021; Park, 2014; Taylor & Barker, 2008). Yet "despite the growing use of corpus linguistic methods in language testing research" (Egbert, 2017, p. 556), a framework for applying corpus data to LTA has not been fully developed. We therefore aimed to make a contribution to this growing literature by taking Peykare's insights into the development and examination of appropriateness and structure of a Persian language proficiency test. To achieve this aim, (a) Persian language features and parameters that may be criterial for distinguishing between proficiency levels were first identified and aligned to the CEFR levels, and (b) validity evidence bearing on constructs (factors) and content domain of the test receptive skills was sought.

To extend prior research in the area of corpora-informed test validation, the present study adopted recent conception of validity as *an integrated evaluative judgment of the degree to which evidential bases support the appropriateness of the test content and*

the interpretation of test scores (see Kane, 2006, 2013; Messick, 1989; Messick, 1993). In this standpoint, content validity per se is insufficient to sustain any testing purpose (Messick, 1993); therefore, content relevance and representativeness should be assessed consistently with generality of the construct interpretation. Given this, the present study adopted a corpus-based approach under an integrated framework of validity to assess both content and construct validity of a Persian language proficiency test designed and intended for measuring PFL learners' communicative language ability.

As to the test content validity, when the assessed parameters were categorized according to the corpus-derived parameters and features, degrees of consistency were observed between the assessed items and the features specified in test specification, providing evidential support for the content relevance of the PLPT-AV. However, two subcategories of POS and Ezafe parameters (i.e., preposition, postposition, and N AJ in Listening section) did not converge with test specification parameters and features proportionally; nor did the categories of homophones/homographs and passive clauses in both Reading and Listening measures. One possible explanation for this can be the "formality" level of contemporary Persian (CP) Standard variety. There exist both formal and informal styles in Standard Persian. Though B2 and C level texts were mostly chosen from academic, education, lectures, art, and architecture registers that are, by nature, formal rather than informal, attempts were made to include more active sentences from daily lives registers in A1, A2, and B1 level texts. In other words, the complexity level of the two parameters was considered in test specifications and designing phases. The reason is that both homophone/homograph meanings and passive structures are more complex for the beginners and B1 intermediates to acquire, thus, they were not targeted for these particular levels.

The focus was not confined to content relevance or coverage alone. It was also on expressions and items conventionally articulated by Persian native speakers, by referring to the frequency of usage or typicality of the items in the corpus. Some correct answers were found to be less typical than distractors, even though they were grammatically correct. According to Pan and Qian (2017) correct answer of infrequent use by native speakers might be considered flawed in content typicality. In addition, because the PLPT-AV is a leveled specific proficiency test, typicality was further examined across the levels. The frequency profiles of A- and B1-level items supported typicality of correct answers, yet atypicality of a few items for these proficiency levels. An important reason for this can be the contextualized tasks of the tests, particularly the text-bound comprehension tasks of Listening and Speaking skills, where controlling parameters could be less feasible than in grammar items, for instance. For this reason, the PLPL-AV receptive sections may not easily include items devoted to assessing specified parameters or their fine-grained categories which fall into grammar assessment which is detached from these comprehension sections of the test. If the test items are decontextualized, assessment of the missing parameters might be done more explicitly. Overall, the expert panel found majority of test items were generally supported by lines of concordances in the corpus, however, in leveled tests like the PLPL-AV if texts and items are selected regardless of typicality profiles, the test might not measure what it is purported to measure at these levels.

In addressing the second research question, i.e., post-administration validity, the construct of the PLPT-AV receptive section was examined in terms of its underlying factor structure. In so doing, the four competing models (i.e., unitary, higher-order, correlated, and uncorrelated) hypothesized based on the available literature were examined through a confirmatory factor analytic approach to see which fit the data better. Each model was tested with data from a sample of PFL learners from different Asian and European countries. The correlated model of distinct listening and reading was identified as the best-fitting model for the test, with the six listening and six reading tasks loading satisfactorily on their associated factors. The regression loads of the model variables appeared moderate to moderately high, whereas the loads estimated for the listening and reading factors in uncorrelated and unitary models were not satisfactory enough to account for an acceptable level of variation explained by the models. The results provide empirical evidence to support the practice of reporting two separate scores corresponding to each section of the test receptive modality. The findings are consistent with the consensus that language ability comprises multiple related but distinct components that are measured by test distinct constructs (see, e.g., In'nami & Koizumi, 2011). The moderately high correlation found between the two factors represent substantial amount of common variance shared by reading and listening, although they are separable skills.

The findings reported for the correlated factor structure of the receptive modalities of the PLPT-AV are congruent and incongruent with the findings reported by other factor-analytic studies conducted on the latent structure of other L2 proficiency tests of the same skills. Although the PLPT-AV is not a test of English language proficiency, the results of its factor structure analyses are consistent with uni-dimensionality for listening and reading comprehension reported by In'nami and Koizumi (2011), nonetheless inconsistent with bi-dimensionality for reading skill and uncorrelated model found by Wilson (2000) in his study on the TOEIC as an L2 proficiency test. The listening and reading factors observed to be correlated in the present study were found uncorrelated by Wilson (2000). This model also differs from the single higher-order factor model reported by Stricker and Rock (2008) to encompass L2 first-order factor. Apart from the language and the content of the PLPT-AV, nature of the samples, and the analytic methods, one possible explanation lies in the designing sources of the tests. The present L2 proficiency test has been structured on Peykare's real language data produced by Persian speakers in real contexts. The parameters used as RLDs across different levels of Persian language ability were purposefully extracted from authentic texts of such a large corpus, i.e., from "written to be read" (associated to reading skill) and "written to be spoken" (associated to speaking/listening) texts. Used for measuring listening and reading abilities, such authentic data might have contributed to producing the uni-dimensionality of these skills. This, in turn, reflects that for each single factor of reading and listening, items are not psychometrically distinct from each other, a finding that is similar to the results of studies on other proficiency tests such as TOEFL test as an international standardized test of proficiency.

The PLPT-AV receptive modules consisted of separate sections/measures that were structured in an increasing difficulty level form using Peykare corpus information. The results of the model testing and validation revealed that the corpus-driven tasks,

accounting for their cumulative contribution to the separate factors of listening and reading, can work appropriately as the distinct indicators of the two factors. Overall, the present results concur with some current reporting of the application of the corpus linguistics to test design and development where the real language tasks contribute to the target construct scores (e.g., Kennedy & Thorp, 2007).

The regression loads of the model variables appeared moderate to moderately high, whereas the loads estimated for the listening and reading factors in uncorrelated and unitary models were not satisfactory enough to account for an acceptable level of variation explained by the models. The results provide empirical evidence to support the practice of reporting two separate scores corresponding to each section of the test receptive modality. The highest regression loads found for certain tasks (e.g., T5 and T6 in Reading; T2 and T6 in Listening) of the confirmed model might be accounted for by the typicality and parameters of the corpus data. Except for Listening T2, a task designed for A-level learners, in C level tasks (i.e., T5 and T6) test developers could include not only the more difficult items but also all types of the targeted parameters with no restrictions. Thus, compared to other items, the C level items are more representative of the domain parameters. On the other hand, item analysis showed typical and the highly frequently used features of natural language for T2 (e.g.,.اتوبوس، کتابخانه، دانشگاه /bus, library, university). These findings lend a degree of support to the significance of including the corpus-driven content in the test.

Precisely speaking, evidence of what language users can do gives a way to the use of such real language data to describe typical abilities common to each proficiency level. The aim of such an approach is to add empirically based linguistic features (e.g., corpus-driven lexical and grammatical details) to functional characterization of the proficiency levels. Therefore, corpora evidence, in particular, when aligned to the targeted CEFR proficiency levels of the tasks, can help test designers to take insights into deciding on specific constructs (i.e., listening and reading) to be tested, writing realistic tasks corresponding to specific proficiency levels, setting realistic criteria to measure what a learner can already do or need to learn in order to achieve mastery of a particular ability level, and validating test constructs and their representative tasks against the real-life texts of various functions.

Conclusions

In this study, we adopted *Peykare* in order to investigate the content and construct validity of a Persian language proficiency test designed for non-native speakers of Persian. Our analysis provided support for the content representativeness, relevance, and typicality of the test. Additionally, the construct validity phase of the study indicated that the correlated model fit the data best in both samples of PFL learners from Asian and European countries.

Taken together, although the present study used Peykare as a Persian database resource for evaluation of a Persian language proficiency test, its findings have implications for LTA, both theoretically and practically. On a theoretical level, the evidence for corpus-informed content validity of such a test signifies the importance of validity evaluation of high-stakes tests against authentic data sources before their final

administration. Such a priori evaluation can reveal certain flaws in covering native speakers' typical expressions by the tests. In practice, the existence of certain unconventionally articulated items in a test introduces sources of invalidity or irrelevant variance to the test, which in turn attenuates its power in assessing test-takers' language knowledge. Therefore, an early evaluation such as content validity, including content relevance, representativeness, and typicality, presents the advantage of using an existing high-quality corpora resource to evaluate and ensure the quality, effectiveness, and purported function of a test. Even though certain flaws are inevitably due to some practical restrictions in item development, care must be taken to cover corpus-extracted language-specific parameters and illustrations that are criterial in determining test-takers' language ability level.

Furthermore, the results shed light on the significance of other corpora-attested evidence about the test: evidence obtained after test administration. This other evidence, in Messick (1993) words, is construct-related evidence. Messick (1989) rejected the traditional notions of validity (i.e., mere content coverage) in favor of an analysis of construct validity which subsumes all other sources of validity evidence. This implies that expert judgments, though made professionally and systematically, might represent test structure inadequately. Based on Kane and Bridgeman (2017) content validity is not enough and needs to be followed by understanding of the corresponding construct.

The evidence for the construct validity of the present corpus-driven test supports the reporting policy and practice of the two separate scores. In fact, a relatively acceptable correlation between the PLPT-AV reading and listening as the two psychometrically distinct factors suggests the argument of separate but related language ability skills. Besides, the application of moment analysis of covariance methodology makes it possible to judge the plausibility of the theoretical model of language proficiency and its components. This has implication for construct validity of large-scale proficiency tests that yield satisfactory sample sizes for further validation studies.

Notably, the present study would provide evidence to promote application of corpora resources in future to other large-scale language tests. From a practical perspective, the application of corpora databases to test material design and development has a washback effect on how real language tasks are designed and how they influence language testing, teaching, and learning. This study demonstrates the usefulness of corpora data for realistic task specification and content not only for testing but for teaching in L2 classes. For instance, the authenticity of the format and content of the PLPT-AV test materials can significantly influence AZFA practitioners' adoption of real-life texts of Persian corpora that are used by the PLPT test designers. Consequently, AZFA teachers would underscore aspects of communicative language abilities underlying the PLPT-AV tasks, which in turn, would involve PFL learners in more learning practices. The findings can also help capture the corpus linguistics state-of-the-art in terms of how it can inform the development and designing of L2 proficiency tests.

However, some limitations of the present study should be noted. First, the sample size in the present study was relatively small. Although the present sample consisted of diverse PFL learners available at the time of the study, it was too small for investigating

the construct in the framework of cross-validation analysis, for instance. When large data sets for the operational PLPT-AV become available in the future, it would be useful to investigate the factor structure within language groups. Therefore, care must be taken in generalizing the present findings to the larger PLPT-AV test-taking population. That is, the findings should be interpreted with caution and a replication should be conducted, for different groups of test takers.

Abbreviations

PFL	Persian as a foreign language
RCISP	Research Center for Intelligent Signal Processing
LTA	Second language testing and assessment

Acknowledgements
Not applicable.

Authors' contributions
MB gave the idea of the work; PS and HM conducted the study, collected the data, and worked on the related literature; the paper is written by PS and edited and revised in different stages by MB and HM. All author(s) read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
The data is available upon request from the corresponding author for using in further research. You may contact the corresponding author.

Declarations

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

Received: 26 December 2022 Accepted: 9 January 2023
Published online: 17 February 2023

References

Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. A. Thomas, & M. H. Short (Eds.), *Using corpora for language research*, (pp. 284–259). London: Longman.
Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383–409.
ALTE. (2002). *The ALTE can do project. Articles and can do statements produced by the members of ALTE 1992-2002.* Retrieved from http://alte.org/downloads/index.php?doctypeid=10.
Arbuckle, J. L. (2009). *IBM SPSS Amos 18 User's Guide*. IBM.
Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly*, *16*, 449–465.
Ball, F. (2001). Using corpora in language testing. In *Research Notes 6, 6-8*. ESOL.
Barker, F. (2004). *Corpora and language assessment: trends and prospects, research notes*. UCLES.
Barker, F. (2010). How can corpora be used in language testing? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics*, (pp. 633–646). Taylor and Francis Press.
Barker, F. (2014). Using corpora to design assessment. In A. J. Kunnan (Ed.), *The companion to language assessment*, (pp. 1013–1028). Wiley-Blackwell.
Batsuren, K., Bella, G., & Giunchiglia, F. (2021). A large and evolving cognate database. *Language Resources and Evaluation*, *56*, 165–189. https://doi.org/10.1007/s10579-021-09544-6.
Beglar, D., & Nation, P. (2013). Assessing vocabulary. *The Companion to Language Assessment*, *2*(10), 72–184.
Bijankhan, M., & ShayesteFar, P. (2016). Corpus-based insights into modeling a level-specific Persian language proficiency test (PLPT): Development and factor structure of the PLPT listening tasks. *Journal of Teaching Persian to Speakers of other Languages*, *5*(*1*), 19–42.
Bijankhan, M., Sheykhzadegan, J., Bahrani, M., & Ghayoomi, M. (2011). Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*, *45*, 143–164. https://doi.org/10.1007/s10579-010-9132-x.
Brooks, L. (2001). Converting an observation checklist for use with the IELTS speaking test. *Research Notes*, *11*, 1–20.
Byrne, B. M. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming*, (2nd ed.,). Lawrence Erlbaum.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: basic concepts, applications, and programming*. Taylor and Francis Group (New York).
Callies, M., & Götz, S. (2015). *Learner corpora in language testing and assessment (Studies in Corpus Linguistics, Band 70)*. John Benjamins.
Council of Europe (2001). *Common European framework of reference for languages: Learning, Teaching, Assessment*. Language Policy Unit.
Council of Europe (2011). *Manual for test development and examining: For use with the CEFR*. Language Policy Division.
Cushing, S. T. (2017). Corpus linguistics in language testing research. *Language Testing*, *34*(4), 441–449.
Egbert, J. (2017). Corpus linguistics and language testing: Navigating uncharted waters. *Language Testing*, *34*(4), 555–564.
Ghonsooli, B. (2010). Development and validation of a PLPT. *Foreign Language Research*, *57*, 115–129.
Gyllstad, H., & Snoder, P. (2021). Exploring learner corpus data for language testing and assessment purposes: The case of verb + noun collocations. In S. Granger (Ed.), *Perspectives on the L2 Phrasicon: The view from learner corpora*, (pp. 49–71). Multilingual Matters. https://doi.org/10.21832/9781788924863-00.
Harrington, D. (2009). *Confirmatory factor analysis*. Oxford University Press.
Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*, 122–159.
Hawkins, J. A., & Filipovic, L. (2012). *Criterial features in L2 English*. CUP.
Hendriks, H. (2008). *Presenting the English profile programme: in search of criterial features, research notes*. UCLES.
Hirai, A. (1999). The relationship between listening and reading rates of Japanese EFL learners. *Modern Language Journal*, *83*, 367–384.
Weir, C., & Milanovic, M. (2003). *Continuity and innovation: Revising the Cambridge Proficiency in English Examination 1913–2002, vol. 15, Studies in Language Testing*. Cambridge University Press.
Haung, L., Kubelec, S., Keng, N., & Hsu, L. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. Language Testing in Asia, 8(14), 1-17. https://doi.org/10.1186/s40468-018-0069-0.
Hughes, A. (2003). *Testing for language teachers*. Cambridge University Press.
Hughes, G. (2008). Text organization features in an FCE reading gapped sentence task. *Research Note*, *31*, 26–31.
Hunston, S. (2022). *Corpora in applied linguistics*. Cambridge University Press.
In'nami, Y., & Koizumi, R. (2011). Factor structure of the revised TOEIC® test: A multiple-sample analysis. *Language Testing*, *29*(1), 131–152.
Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper (Research Memorandum No. RM-00-03)*. ETS.
Kane, M. (2006). Validation. In R. Bernnan (Ed.), *Educational measurement*, (4th ed., pp. 17–64). American Council on Education and Praeger.
Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73.
Kane, M., & Bridgeman, B. (2017). Research on validity theory and practice at ETS. In R. Bennett, & M. von Davier (Eds.), *Advancing human assessment: The methodological, psychological and policy contributions of ETS*, (pp. 489–552). Springer Open. https://doi.org/10.1007/978-3-319-58689-2.
Karimi, Y. (2007). Kurdish *Ezafe* construction: Implications for DP structure. *Lingua*, *117*(12), 2159–2177. https://doi.org/10.1016/j.lingua.2007.02.010.
Kennedy, C., & Thorp, D. (2007). A corpus-based investigation of linguistic responses to an IELTS Academic Writing Task. In L. Taylor, & P. Falvey (Eds.), *IELTS Collected Papers: Research in Speaking and Writing Assessment (Studies in Language Testing vol. 19)*, (pp. 316–377). UCLES and Cambridge University Press.
Kline, R. B. (2011). *Principles and practice of structural equation modeling*, (3rd ed.,). The Guilford Press.
MacCallum, R. C., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research*, *29*, 1–32.
McCarthy, M. (2010). Spoken fluency revisited. *English Profile Journal*, *1*(*e4*), 24–39. https://doi.org/10.1017/S2041536210000012.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). Macmillan.
Messick, S. (1993). Foundations of Validity: Meaning and consequences in psychological assessment. *ETS Research Report Series*, *2*, i–18. https://doi.org/10.1002/j.2333-8504.1993.tb01562.x.
Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256.
North, B. (2014). *The CEFR in practice*. Cambridge University Press.
Oller Jr., J. W. (1983). Evidence for a general language proficiency factor: An expectancy grammar. In J. W. Oller Jr. (Ed.), *Issues in language testing research*, (pp. 3–10). Newbury House.
Pan, M., & Qian, D. D. (2017). Embedding corpora into the content validation of the grammar test of the National Matriculation English Test (NMET) in China. *Language Assessment Quarterly*, *14*(2), 120–139. https://doi.org/10.1080/15434303.2017.1303703.
Park, K. (2014). Corpora and Language Assessment: The State of the Art. *Language Assessment Quarterly*, *11*, 27–44.
Purpura, J. E. (2004). *Assessing Grammar*. Cambridge University Press.
Römer, U. (2022). Applied corpus linguistics for language acquisition, pedagogy, and beyond. *Language Teaching*, *55*(2), 233–244.
Salamoura, A., & Saville, N. (2009). Criterial features of English across the CEFR levels: evidence from the English Profile Program. *Research Notes*, *37*, 34–40.
Taylor, L., & Jones, N. (2006). Cambridge ESOL exams and the Common European Framework of Reference (CEFR). *Research Notes*, *24*, 2–5.
Sang, F., Schmitz, B., Vollmer, H. J., Baumert, J., & Roeder, P. M. (1986). Models of second language competence: A structural equation approach. *Language Testing*, *3*(1), 54–79.
Wilson, K. M. (2000). *An exploratory dimensionality assessment of the TOEIC test*, *TOEIC Research Report, RR-00-14* (). Educational Testing Service.
Saville, N. (2003). The process of test development and revision within UCLES EFL. In C. J. Weir & M. Milanovic (Eds.), *Continuity and innovation: revising the Cambridge proficiency in English examination, 1913-2002* (pp. 57–120). Cambridge University Press.

Saville, N., & Hawkey, R. (2010). The English Profile Programme: the first three years. *English Profile Journal*, *1*, e7. https://doi.org/10.1017/S2041536210000061.
Simpson, R. C., Lee, D. W., & Leicher, S. (2002). *MICASE manual. The Michigan Corpus of Academic English*. The University of Michigan.
Stricker, L. J., & Rock, D. A. (2008). *Factor structure of the TOEFL internet-based test across subgroups. TOEFL iBT. Research Report, RR-08-66*. ETS.
Taylor, L., & Barker, F. (2008). Using corpora for language assessment. In E. Shohamy, & N. H. Hornberger (Eds.), *Encyclopedia of Language and Education*, (vol. 7, 2nd ed., pp. 241–254). Language Testing and Assessment. Springer.

Publisher's Note