Language Testing in Focus: An International Journal

Language Testing in Focus An International Journal LTiF

ISSN: 2717-9087

www.EUROKD.com

2025 (11)

Poetry-based cloze tests as a language assessment tool for European Portuguese as a foreign language

Clara Setas[1*], Beatriz Oliveira[2]
[1]*Universidade do Minho, Center for Humanistic Studies of the University of Minho, Portugal*
[2]*Universidade de Aveiro, Center for Languages, Literatures and Cultures, Portugal*

ABSTRACT

This study explores the use of poetry-based cloze tests as a language assessment tool for European Portuguese (EP) as a Foreign Language (PFL) at beginner-to-intermediate proficiency levels. Three rational cloze tests, developed from Portuguese poetry, were pretested with 24 native speakers and validated with 32 PFL learners. Data included response accuracy, perceived test difficulty, and sociolinguistic background. Test- and item-level analyses confirmed high internal consistency and item discrimination for the intended construct. Learners' sociolinguistic profiles – self-assessed proficiency, country of residence, and native language – influenced performance. Item design features, namely hints and the structured response format for omitted words, may have contributed to the consistency in complexity across all cloze tests. This study highlights the role of poem selection criteria in test development, such as cultural relevance, narrative coherence, and lexical richness. Findings support the integration of poetry into PFL testing practices to promote holistic test development. These findings provide a methodological contribution to the field of PFL assessment, encourage the comparability of cloze testing methods in PFL research involving similar learner populations.

Keywords
Poetry, Language Testing and Assessment, Cloze Test, Validity Evidence, Portuguese as a Foreign Language

Received
15 April 2025
Received in revised form
14 September 2025
Accepted
12 October 2025

Correspondence concerning this article should be addressed to: id10885@uminho.pt

How to cite this article (APA 7th Edition):
Setas, C., & Oliveira, B. (2025). Poetry-based cloze tests as a language assessment tool for European Portuguese as a foreign language. *Language Testing in Focus: An International Journal*, *11*, 55-82. https://doi.org/10.32038/ltf.2025.11.04

Introduction

Introducing poetry in foreign language classrooms requires careful selection and appropriate exploration of literary texts, and a dedicated effort from the educators who might face difficulties, in practice, due to the scarce prescriptive and guiding curriculum (Sánchez, 2017). The true shift in foreign language learning approaches occurred in the 1980s with the communicative approach. During this period, literature gradually became incorporated into the foreign language teaching and learning process paradoxically, as asserted by Plaza (2009). For instance, Bachman and Palmer (1996) emphasize the importance of aligning test tasks with real language use to enhance authenticity and validity. As literary language serves as a reflection of a society's way of thinking and living, poetry can be an essential resource for understanding the culture and way of thinking of speakers of a target language (Oliveira, 2018). Poetry exposes students to authentic language that can develop their language ability (Reazul, 2022).

Language used in poetry is not devoid of communicative character or disconnected from reality, it represents a type of knowledge that goes beyond simple communication, evoking memory, metaphors, and the ineffable as special ways to capture reality (Ribeiro, 2007). Concerning its language, poetry plays with grammar and syntax in creative ways and frequently employs rich, diverse, and descriptive vocabulary, as well as idiomatic expressions, and poetic devices, such as metaphor, simile, alliteration, and imagery (Mermelstein, 2022). Additionally, poetry showcases “complex treatments of language… [and a] focus on "how it is said" as [much as] "what is said"' (Weaven & Clark, 2013, p. 198). According to Férez Mora and Coyle (2020), students’ motivation improved due to the poem’s authenticity, briefness, and the non-triviality of its topic. Given its favorable characteristics, it is surprising that very few studies focus on poetry in Portuguese as a Foreign Language.

Beyond its potential in foreign language education, poetry offers a richness that surpasses cultural and linguistic dimensions. Poetry-based texts possess unique qualities that make them conducive to assessment practices, especially, learning-oriented assessments (LOA). Poetry's dense language, nuanced structures, and layered meanings benefit the assessment of learners' comprehension, language ability, and critical thinking abilities (see Peskin, 2007; Wright et al., 2010). Despite the positive outcomes underlined, poetry is scarcely used in formative and summative assessment instruments. It is rather mostly applied within references of target-culture teaching and learning, in listening and writing comprehension exercises, concerning PFL teaching. Moreover, no studies to date have begun to explore poetry-based tools for PFL assessment, therefore the present study aims to fill this gap in PFL assessment practices.

Literature Review

Poetry in Foreign Language Teaching and Testing

Poetry holds strong pedagogical value due to its intrinsic and relational characteristics, functioning as an authentic reality and a channel for meaningful communication (Ribeiro, 2007). Its metaphorical density stimulates imagination and creativity, establishing a connection between poetry and knowledge. Research shows its benefits in improving productive and receptive skills (Lee & Lin, 2015; Mittal, 2014). Poetry's rhythmic and repetitive nature aids in memorizing lexical items and syntactic structures (Fabb, 2015), which supports faster

vocabulary retention. Vocabulary is fundamental in language acquisition (Meara, 1996), and its structured development is crucial (Nation, 2002), as it should be organized, and regularly assessed. Thus, poetry can be used to better assess vocabulary, grammar and pronunciation, especially in teenagers rather than younger learners (Kanonidou & Papachristou, 2019).

The use of poetry must be dependent on the target-audience, despite its multidimensional nature. Martínez (2008) notes that poetry often deviates from standard grammatical conventions, making it challenging for learners focused on communicative skills. The elaborate language and literary complexity can overwhelm students, reinforcing the need for careful poem selection for lower proficiency levels. In fact, in the PFL reference framework *Referencial Camões PLE* (Camões, I.P., 2017), poetry (specifically, simple and brief) as a genre only appears in the private oral domain for B2 level of proficiency. This implies that the use of poetry in-class is rarely applied and incentivized in PFL classrooms, internationally. However, poetry is rich in content, vocabulary in context (Kellem, 2009), and language use, therefore it can be implemented across levels and age groups (Kanonidou & Papachristou, 2019). Schander et al. (2013) argue that poetry can be tailored to various complexity levels. A systematic review and recent analyses of LOA (Wakid, 2024) showed that authentic tasks, actionable feedback, and learner involvement are key elements, positioning poem-related tasks as a strong LOA fit. Given its demonstrated effectiveness as a pedagogical tool (Ribeiro, 2007), poetry merits integration not only into regular in-class language assessments, but also more optimally, within the LOA framework.

Cloze Tests for Language Assessment

Across diverse linguistic contexts, language assessments are indispensable, and cloze tests are especially effective for their capacity to assess vocabulary knowledge and reading comprehension. Vocabulary is the core of L2 proficiency and studies have shown a high correlation between proficiency and the measure of productive vocabulary (Alderson, 1979; Purpura, 1999). Cloze tests, as a fill-in-the-gap format, evaluate morphosyntactic structures, targeting grammatical accuracy and lexical awareness, requiring test-takers to reconstruct passages using all their language knowledge, as integrative assessments (Chung & Ahn, 2019). In the context of the present study, lexical awareness involves not only recognizing grammatical structures but also demonstrating an understanding of how words function in a poetic text – their meaning, nuances, and appropriate usage. Usually, words are fully omitted within cloze tests (for example in Tremblay & Garrison, 2010), however more recent studies have started to also partially omit words (for example, Flores et al., 2021; Flores et al., 2022) to control test construct and item response accuracy. Even if words are not fully but partially omitted, test takers need to understand meaning and complete the words, fitting the context both semantically and stylistically. Research in L2 sentence processing (e.g., VanPatten, 2015) suggests that learners rely on lexical meaning before attending to grammatical details.

Cloze tests have been extensively studied in L1 and L2 contexts, supporting their effectiveness in proficiency assessment (Oller & Jonz, 1994; Watanabe & Koyama, 2008). However, opinions on their reliability vary across proficiency levels (Alderson, 1979; Fotos, 1991). Performance on these tests are influenced by contextual factors, including cultural background

and language exposure (Rahimi, 2014), which signifies they should be carefully designed for a target-audience. Despite its challenges, cloze tests remain a structured, controlled assessment method for categorizing speakers and evaluating diverse language structures (Rinke et al., 2024). Furthermore, studies have proven cloze tests are internally consistent, even across different cloze test formats (e.g., Bachman, 1985; Chapelle & Abraham, 1990).

Test design requires careful consideration of material selection, item characteristics, response types, and deletion rates (Park, 2011). Recent work by Hossain (2024) shows that integrating poetry boosts motivation and productive skills when paired with careful text selection and scaffolding. Omitted words may be represented by lines or dashes with in-text responses, which have been proven to generate higher scores than separate response boxes (Hartley & Trueman, 1986). A balanced distribution of linguistic complexity is recommended, across approximately 50 cloze gaps (Oller, 1979; Flores et al., 2022). The effectiveness of cloze tests in L2 assessment is well-documented (Tremblay, 2011), nevertheless further research is needed to assess their reliability and validity in PFL and different speaker populations (Kalyoncu & Memiş, 2025).

Complexities of Language Assessment

A survey conducted by Tremblay and Garrison (2010) demonstrated the lack of uniformity among researchers in deciding which proficiency assessment method to apply and in characterizing proficiency tools. Despite this, language assessment research continues to refine methods and tools that positively influence testing practices. Researchers should focus more on ethical responsibilities (McNamara, 2001), fairness (Kunnan, 2004), societal impact (Roever & McNamara, 2006), and instructional application (Wall, 2005) of language assessments. However, the assessment development process becomes heavily reliant on the validation and reliability analysis, both critical for the quality of assessment use. The validation process itself is fraught with challenges, one being the scarcity of comparable tests for reference, in the efforts to establish reliability. Lousada et al. (2012) discussed the difficulty of establishing concurrent validity without validated benchmarks. There is a need for more validated tools across languages and domains. While the choice of test content and design becomes more goal- and audience-oriented, the work of Schmitt et al. (2020) in turn argues the necessity for more rigorous validation procedures. Validation is not a static quality, but a dynamic measure reflecting the relationship between test scores and learners' competence (Le & Klein, 2002). Thus, scoring procedures should be clearly specified, empirically refined, and aligned with the intended construct and language characteristics of the target population (see Effatpanah et al., 2025; Kalyoncu & Memiş, 2025).

Language descriptors are critical to the assessment development and score interpretation processes, particularly in educational contexts. However, the abstract nature of language makes assessment difficult (Iskandarova, 2024). Studies suggest significant variations in language complexity (Ehret & Szmrecsanyi, 2016), and frameworks such as the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2020) may not fully align with local mandates, particularly in linguistically diverse contexts (Elatia, 2011) – such as in Portugal, with its immigrant population and regional language varieties.

Research Questions

The following research questions were formulated:

RQ$_1$: *Can poetry-based cloze tests be used as a viable language assessment tool for EP as a foreign language?*

The aim is to find validity evidence in three poetry-based cloze tests designed for an audience of (young) adult PFL learners, using relevant test- and item-level statistics. We do not intend to promote cloze tests as the sole effective method for assessing proficiency in foreign language or L2 research and teaching, as they may not be suitable for all learner populations and are insufficient for evaluating, for example speaking skills. Rather, we emphasize the importance of incorporating multiple authentic assessments to ensure an accurate interpretation of proficiency, particularly in certification contexts.

It is anticipated that the tests will demonstrate overall moderate to high validity evidence and good discrimination power across test-takers of different proficiency levels of PFL.

RQ$_2$: Can the target linguistic structure groups being assessed predict participants overall performance separately?

The aim is to identify differences in complexity across groups of target linguistic structures that might explain or even predict overall performance.

It is expected that the groups of linguistic structures will contribute differently to participants' overall performance, with more complex structure groups posing greater challenges and potentially leading to lower accuracy rates. If a strong correlation is found between specific structure groups and overall performance, this would suggest that those serve as valid indicators of proficiency.

RQ$_3$: Is there a correlation between test-taker' scores and their sociolinguistic profile?

The aim is to determine input factors of influence in the test-takers' language acquisition process to establish possible correlations – within the small sample of participants.

A correlation between test-taker' scores and some characteristics of their sociolinguistic profiles is expected. Specifically, concerning factors such as country of residence, number of years spent learning European Portuguese and weekly hours dedicated to learning are likely to influence scores. Additionally, self-assessment reports are expected to predict test-taker's performance in the cloze tests, both individually and altogether.

Methodology

Participants

Two participant groups were recruited and assigned to two different phases: native EP speakers for the pretesting phase, and PFL learners for the validation phase. Both groups completed a sociolinguistic questionnaire and three cloze tests assessing EP language proficiency. The pretesting phase included 24 native Portuguese speakers (ages 19–32, M = 26.25, SD = 3.44),

all born and residing in Portugal. As for the validation process with PFL speakers, the participants (32 in total) represent a diverse range of nationalities. Most participants are young adults (19-23 years), with a sparse representation of older individuals plus 36 years old (mean = 24.25, SD = 7.80). In this group there were 27 females (84.4%) and 5 males (15.6%). The largest subgroup was Croatian (34.4%), followed by Turkish, Spanish, and Slovenian (each 9.4%). Austrian participants made up 6.25%, while the remaining 3.1% represented each of the following nationalities, Polish, Serbian, Japanese, Chinese, Mongolian, Ukrainian, Iranian, Colombian and Congolese. One participant did not disclose nationality. Regarding native languages, Croatian speakers comprised 34.4%, followed by Slovenian and Spanish (12.5% each), Turkish (9.4%), and German (6.3%). The other languages, French, Japanese, Mandarin, Persian, Mongolian, Polish, Serbian, and Russian, accounted for 3.1% each. Croatia and Portugal had the highest country of residence representation (34.4% each), followed by Slovenia (15.6%) and Spain (9.4%). Poland and Austria each accounted for 3.1%. Most participants have limited years of learning EP (15 participants report only 1 year of learning), except for a single outlier with 20 years of exposure. All participants, except three (9.4%), have provided self-assessment for PFL: 6 participants self-assessed as beginner (18.8%), another 6 as beginner-intermediate (18.8%), 11 as intermediate (34.4%), and 5 as intermediate-advanced (15.6%). We did not collect data on how long participants have been living in the country of residence. Table 1 summarizes participant age, years of learning EP, and weekly study hours.

Table 1

Descriptive Statistics for Age, Learning Duration, and Weekly Learning Hours

	Minimum	Maximum	Mean (M)	Standard Deviation (SD)
Age (years)	19	53	24.25	7.80
Years of learning EP	0	20	2.29	5.34
Weekly EP learning hours	3	24	8.46	4.67

Data Collection Methods

At both phases, participants completed three rational cloze tests sequentially, with data collected on response accuracy, perceived difficulty, and sociolinguistic profiles. The study was conducted online via *Microsoft Forms* and *Cognition.run*. The testing environment was not controlled. The sociolinguistic questionnaire gathered data on age, gender, nationality, native language(s), and learning history, including years and weekly hours spent studying EP, and self-assessment. The three cloze tests were based on original Portuguese poetry: *Um Livro* (Mésseder, 2003), *Perdi os Meus Fantásticos Castelos* (Espanca, 2012), and *Imaginação* (Lobato de Faria, 1996). Items were omitted as full or partial words, within rectangular input boxes, which also indicate missing letter counts for each item - except in Test 3, where no clues regarding the correct response were provided. Participants first completed a brief training exercise to familiarize themselves with the format. Responses were categorized into four coding options: correct, incorrect, missing, or unexpected but correct. For statistical analysis purposes, correct and unexpected but correct responses were coded as (1), while incorrect and missing responses were coded as (0). Unexpected correct responses were reviewed by the authors, who are both native EP speakers. Spelling errors were not considered incorrect. These

cloze tests were built to assess the proficiency of (young) adults learning PFL in a formal context at beginner to intermediate levels.

Corpus Selection

Researchers along the years have identified that cloze tests are influenced by text complexity (Bachman, 1985; Yamashita, 2003). Effective cloze test design requires unbiased, level-appropriate, self-contained passages of suitable length (Oller, 1979). Following Vardell et al. (2011), the selected poems exhibit linguistic richness, thematic diversity, and clear comprehension flow, making them ideal for PFL assessment. The three poems address accessible universal themes, namely reading as a journey, loss and disillusionment, and imagination. Their brevity allows completion within 15 minutes. This aligns with Gardihewa (2022), who found that simple, meaningful poems can enhance learners' comprehension and language skill development in the target language.

Item Selection

The identification of item complexity facilitates the selection of the item pool for a rational cloze test aiming to integrate more than one complexity level. This study's item selection (28 items) stems from the existing literature on linguistic complexity of EP structures from A1 to B2 of the CEFR levels and is oriented by the framework in *Referencial Camões PL*E[1]. The linguistic structures were selected within the available lexical richness of each poem, striving for a balance in complexity levels along the texts. To facilitate statistical analysis, the structures were grouped by categories (see Appendix 1) with the following distribution: Group 1 (G1_V_INFL) comprises verbs and inflected verbs; Group 2 (G2_PRO_R_P_I) consists of relative, possessive, and interrogative pronouns; Group 3 (G3_PREP_CONTR) integrates prepositions and contracted prepositions; Group 4 (G4_DET_N_MODIF) includes determiners, nouns, and modifiers; and Group 5 (G5_PRO_CLIT_POLA) covers clitic pronouns and polarity operators.

Regarding item complexity, highly challenging structures (primarily from formal Portuguese registers requiring regular exposure for accurate acquisition), include clitic pronouns ([-lo], [-lhe], [-a]), the relative pronoun *que*, and contracted prepositions (*pelo*). Other complex structures include the interrogative pronoun [*qual*], the preposition [*com*] and the conjunction [*que*]. These structures require regular exposure to written registers for accurate acquisition and use (Flores et al., 2022). Specifically, for the acquisition of clitic pronouns Fiéis et al. (2023) also argue that the full acquisition of patterns of clitic positioning is possible but only at a native-like level. Hence, language classes can provide consistent exposure (quantity of input) to diverse registers of the Portuguese language, exploring language use through different types of resources (spoken and written formats), while also enhancing students' metalinguistic knowledge of the language. As argued in Rinke et al. (2024), complexity in acquisition is a multifaceted phenomenon influenced by various linguistic and cognitive factors. Based on

[1] The *Referencial Camões PLE* is an educational framework developed by Camões, I.P., aligned with the CEFR, providing level descriptors (A1–C2) and inventories of pragmatic, notional, and linguistic content to guide the teaching, learning, and assessment of PFL.

previous studies on complexity in monolingual acquisition (Costa et al., 2015; Charneca Catalão, 2011; Batalha, 2018), the authors identified four types of complexity: derivational complexity, which involves syntactic operations such as embedding and movement (e.g., relative clauses; Armon-Lotem, 2005; Costa et al., 2011; Vasconcelos, 1995); irregular and lexical forms, which rely on memory rather than rule-based processing (e.g., verb-preposition selection); context-dependent rules, which require the integration of syntax, discourse, and phonology (e.g., clitic allomorphy; Costa & Lobo, 2009; Costa et al., 2012; Flores et al., 2020); and multiples form-function mappings, where a single form serves different grammatical functions (e.g., [*que*], [*por*], and the inflected infinitive in concessive clauses; Santos, 2017; Costa, 2006). The item selection process for the present study followed the guidance of these findings.

Appendix 1 provides the selected poems and items, along with passage context, structure description, and structure identification used in the statistical analyses. The most complex structures include clitic and relative pronouns, contracted prepositions, and inflected infinitives. Other complex structures (but less than the previous ones) comprise possessive pronouns, verb inflections, and contracted prepositions, covering also interrogative pronouns, adjective inflections, adverbs, and relative pronouns. As for the easiest structures, verb inflections, articles, and noun inflections were included. The relative pronouns appear in complex and easy structures due to differences in syntactic embedding. This study aimed at finding validity evidence in these items considering the present construct.

Findings and Discussion

Pretesting of the Cloze Tests with Native Speakers of EP

From 24 native speakers, 602 responses were correct (89.58%) and 70 incorrect (10.41%). Most incorrect responses occurred in Poem 1 (55.71%), followed by Poem 2 (27.14%) and Poem 3 (17.14%). The mean score (25.08, SD = 2.90) indicates high performance with moderate variability (Table 2). Reliability analysis confirmed internal consistency ($\alpha = 0.73$; Guttman's L4 = 0.95).

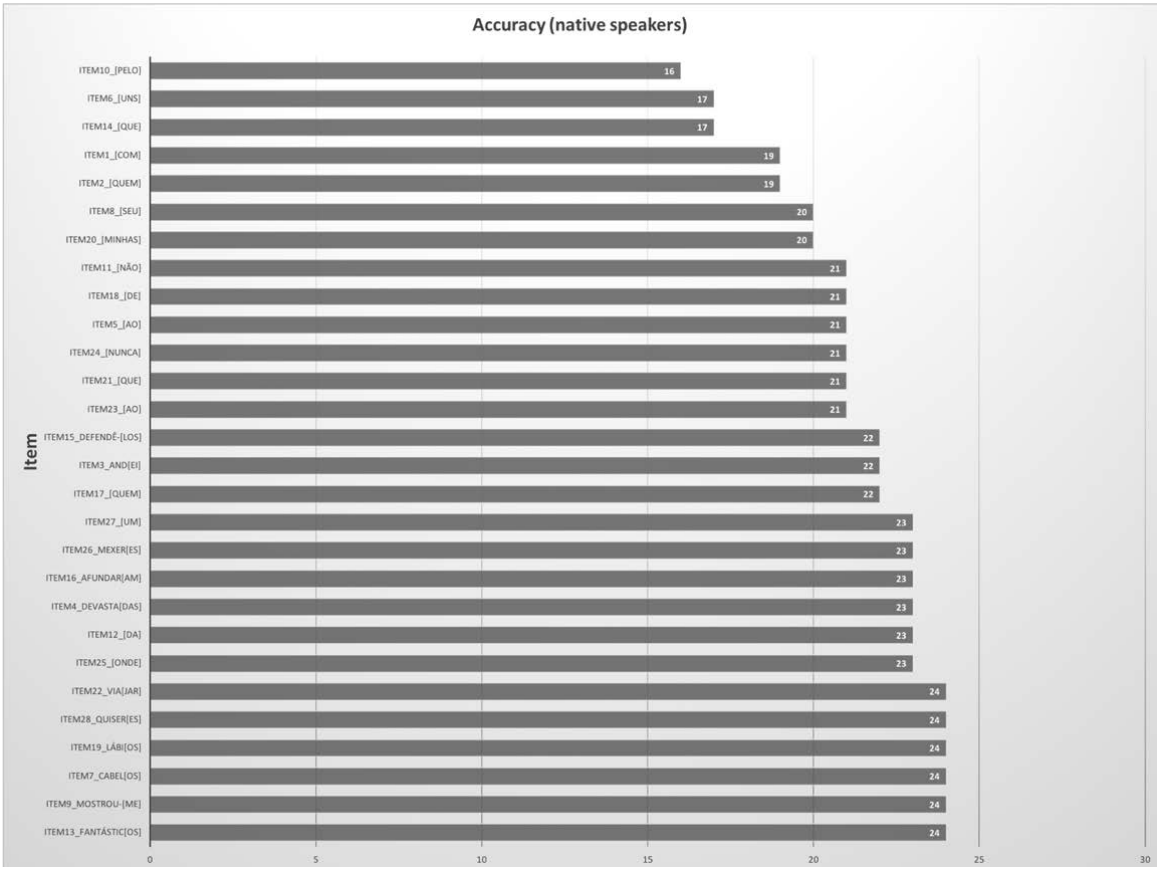
Table 2

Overall Results of the Cloze Test for the Group of Native Speakers

	Correct responses	Mean	SD	Min-Max.	Cronbach's Alpha	Guttman's L4
Native speakers	602	25.08	2.90	16-28	0.73	0.95

Figure 1

Accuracy of Item Responses from Native Speakers



The most frequent incorrect responses as shown in Figure 1 occured in the following items:

a. The preposition '*por*' contracted with the definite article '*o*', [*pelo*] (Test 1, Item10).
b. The relative pronoun [*que*] in a subordinate adverbial clause of manner (Test 2, Item14).
c. The indefinite article [*uns*] in the masculine and plural form (Test 1, Item6).
d. The relative pronoun [*quem*] (Test 1, Item2).
e. The preposition [*com*] (Test 1, Item1).

These results align with previously mentioned complexity findings. It is important to note that, in addition to incorrect responses, certain alternative responses were accepted for specific items. In Test 1, [*sem*] was accepted in place of [*com*] (a direct antonym), and [*nem*] was accepted instead of [*não*] (a synonym and an adverb of negation). Additionally, [*de*] was accepted without its contraction with the definite article [*a*], as this form is grammatically correct and dependent on stylistic choice. In Test 2, [*como*] and [*onde*] were accepted as alternatives to [*quem*], since these also function as interrogative pronouns. Lastly, in Test 3, [*não*] was accepted in place of [*nunca*], which both serve as adverbs of negation. Overall, only a few items across the cloze tests elicited one or two alternative correct responses.

PFL Learner's Results at Test- and Item-Level

After the pretesting, the tests were administered to PFL learners. A Cronbach's alpha reliability analysis was conducted on the items of the cloze tests altogether. This analysis revealed

excellent internal consistency (α = 0.93), indicating that the items measure and contribute to a similar construct (see Table 3). The Kuder–Richardson reliability coefficient (KR-20) for the tests altogether was 0.94, which reinforces the internal consistency results. KR-20 is another measure of internal consistency specifically designed for dichotomous items. This indicates that the items function coherently in measuring the same or similar underlying construct and that the observed test scores are highly stable and precise. While such a high level of reliability is desirable, it may also reflect a degree of item redundancy, entailing that several items could be providing overlapping information rather than contributing unique variance. From a psychometric perspective, this reinforces the test's precision but may narrow its capacity to capture a wider range of the construct at its extreme points, namely very low or very high proficiency. In this case, this does not constitute a major concern since the target is set at discriminating among beginner to intermediate proficiency levels. Figure 2 reports the number of correct responses per item combining all tests (from more complex to less complex). We did not run a reliability analysis for each test individually since it is not recommended due to the small number of items per test. Among the three tests, Test 2 had the highest average score (0.684; SD = 0.466), followed closely by Test 3 (0.672; SD = 0.470) and Test 1 (0.609; SD = 0.489), which suggests Test 1 may have been slightly more challenging for the learners. It is important to note that the overall omission rate was very low (5,6% in total for the tests altogether).

Table 3

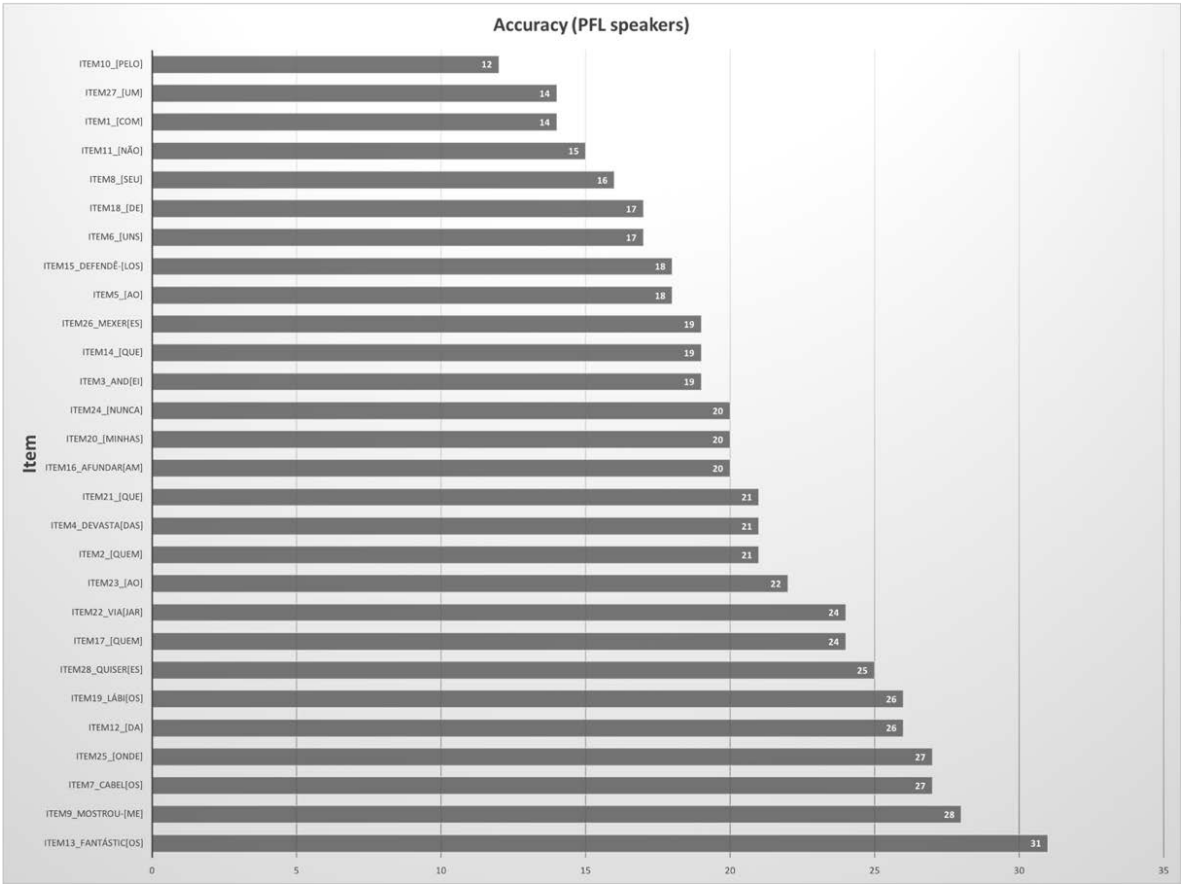
Overall Results of the Cloze Tests Altogether for the Group of PFL Speakers

	Correct responses	Mean	SD	Min-Max.	Cronbach's Alpha	Guttman's L4
PFL speakers	581	18.15	7.68	2-28	0.93	0.99

A one-way ANOVA analysis was conducted, to assess whether the three cloze tests differed from each other regarding test performance and response accuracy. The results indicate that the difference in test performance across tests was not statistically significant (p = 0.102). This confirms that the difficulty levels across tests are comparable. Guttman's L4 analysis is 0.99, which also shows very high reliability. The mean test score is 18.15 (out of 28), with an SD of 7.68. This shows a relatively wide spread of scores (see Figure 3 for participant total scores). There is a notable ceiling effect: many participants scored close to the maximum (for example, 23 participants scored above 50%).

Figure 2

Accuracy of Item Response from PFL Speakers



The best-performance items based on PFL learners' results were:

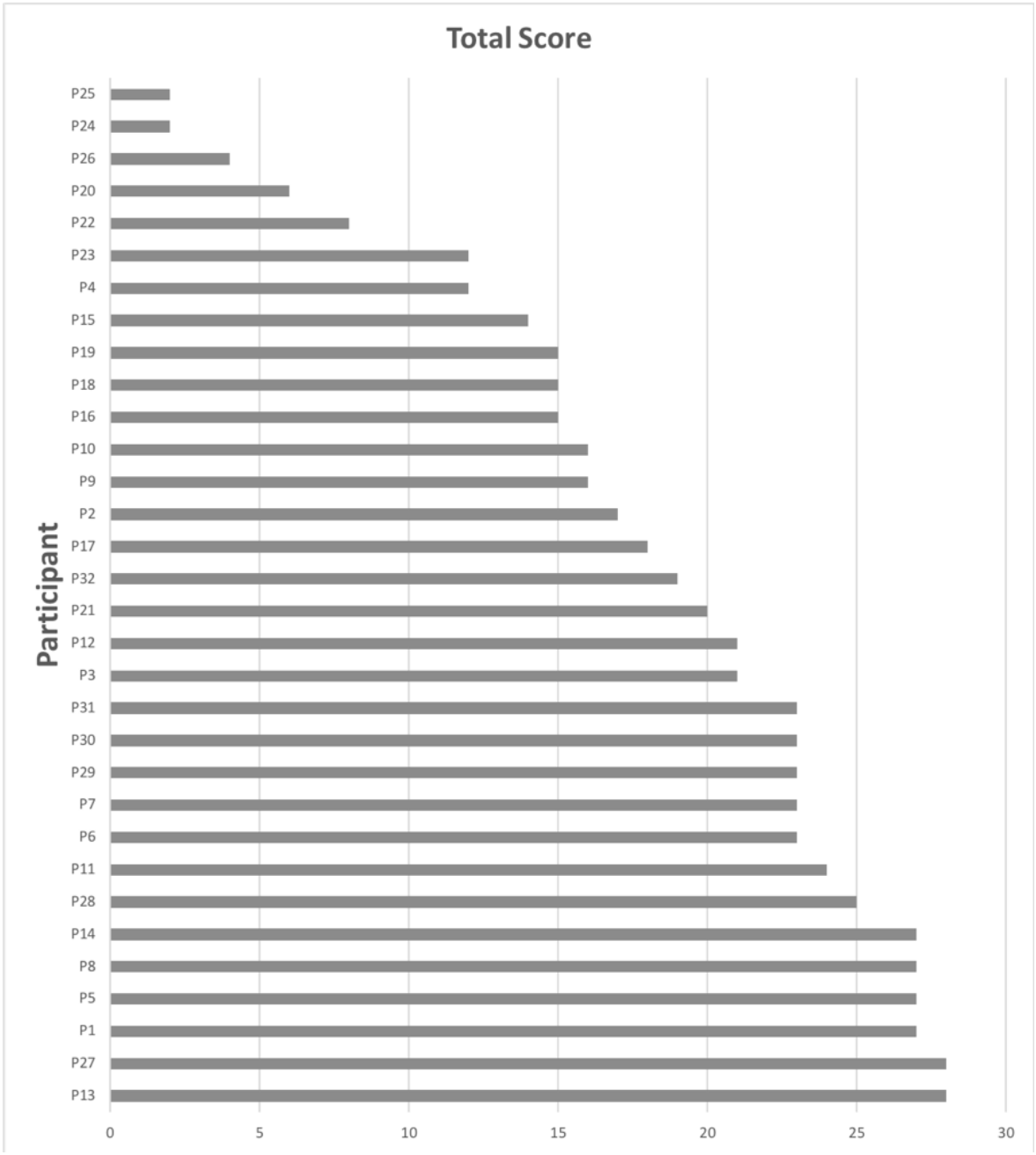
a. The clitic pronoun in the dative form *mostrou*[*-me*] (Test 1, Item9).
b. The adjective inflection *fantástic*[*os*] in the masculine, plural form (Test 2, Item13).
c. The relative pronoun [*onde*] in an adverbial subordinate clause of place (Test 3, Item25).

Conversely, the most frequent incorrect responses occurred in:

d. The contracted preposition [*pelo*] in masculine, singular form (Test 1, Item10).
e. The preposition [*de*] in simple form (Test 2, Item18).
f. The indefinite article [*um*] in masculine, singular form (Test 3, Item27).

Figure 3

Participant Distribution by Total Score



To further detail the analysis of item performance and understand test taker characteristics, both Classical Test Theory (CTT) indices and Item Response Theory (IRT) - using a Two-Parameter Logistic Model (2PL) - were analyzed (see Table 4). Item facilities ranged from 43.8% (Item1 and Item27) to 65.6% (Item2 and Item4), with most items clustering around intermediate facility levels. The very easy items were identified as, Item7, Item9, Item12, Item13, Item19, Item25, and Item28. The easy items were Item17 and Item22, while the rest of the items ranged between moderately easy to moderately difficult. Corrected item-total correlations (R_rest) were consistently strong across the tests altogether, exceeding 0.60 for most items which indicates good discrimination power between high- and low-performing participants across the sample's ability range (beginner-intermediate). For example, Item2 and

Item5 were among the best-performing items, effectively differentiating higher from lower scorers. A few items demonstrated lower but still acceptable discrimination, such as Item4, Item8, Item9, Item13, and Item22, representing weaker alignment with the overall construct. The IRT parameter estimates (Table 4) provided further insights into item performance. Item discrimination values (*a* parameters) varied considerably, with some items displaying weak discrimination slopes (e.g., Item4, $a = 0.45$) and others reflecting strong sensitivity to ability differences (e.g., Item5, $a = 4.50$; Item18, $a = 12.48$). These results show that while most items align well with the latent construct, a subset demonstrates particularly high discrimination. Item difficulty parameters (*b* values) covered a fair range, from easier items such as Item9 ($b = -3.72$) and Item13 ($b = -3.69$) to more challenging ones such as Item10 ($b = 0.67$) and Item27 ($b = 0.33$). This confirms that the test best captures abilities in the lower-to-intermediate range, with limited coverage of higher ability levels.

Table 4

Classical Test Theory and Item Response Theory Analysis (Based on PFL Participants' Results on all the Items)

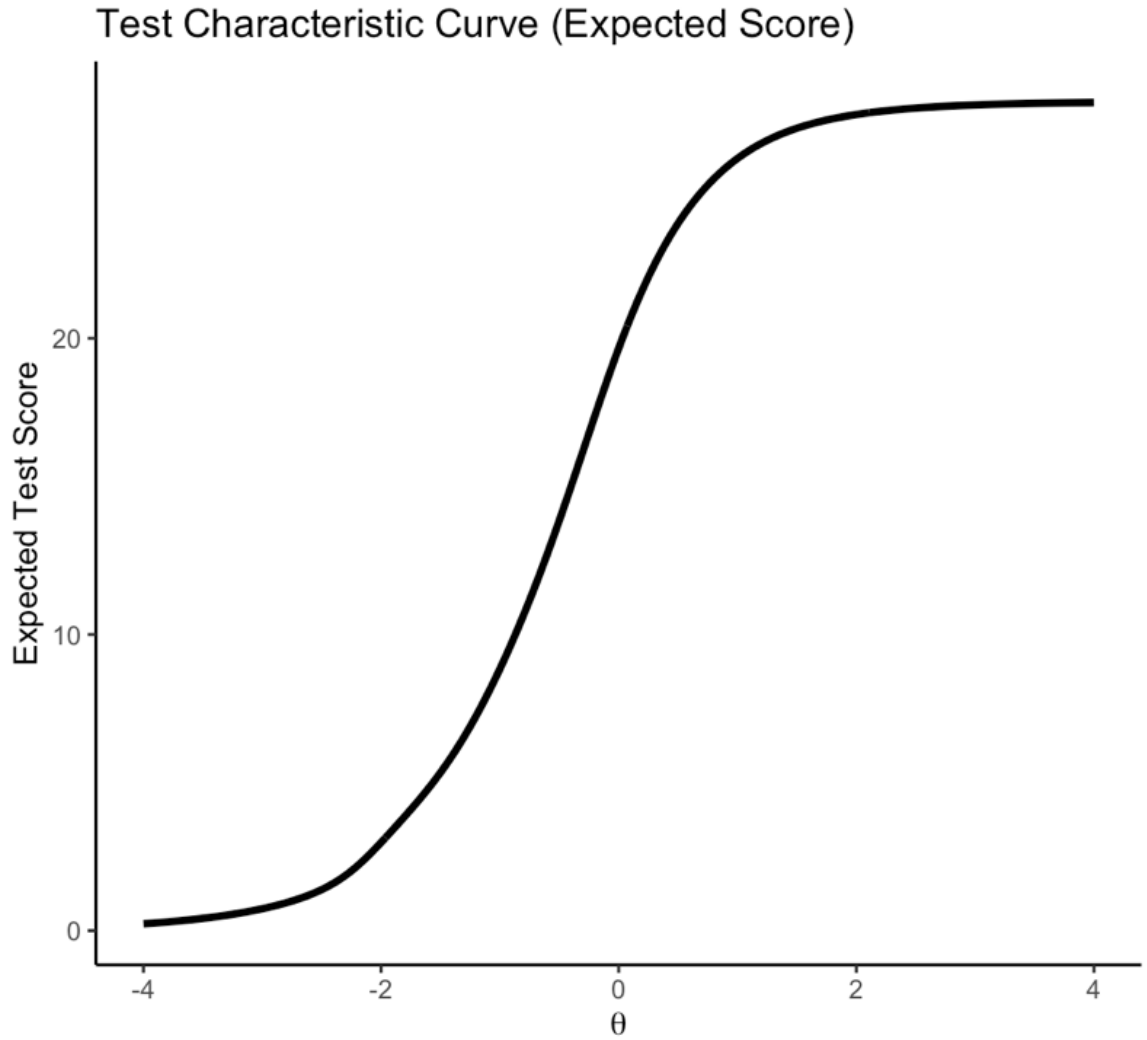
	CTT				IRT					
	Mean	SD	Facility	R_rest	*a*	*b*	Outfit	z.Outfit	Infit	z.Infit
Item1	0.44	0.5	43.8	0.60	2.40	0.10	0.61	-0.35	0.86	-0.61
Item2	0.66	0.48	65.6	0.71	4.12	-1.08	0.54	-0.16	0.78	-0.64
Item3	0.59	0.5	59.4	0.42	1.05	-2.07	0.89	-0.44	0.98	-0.09
Item4	0.66	0.48	65.6	0.38	0.96	-0.29	1.21	0.88	0.98	-0.10
Item5	0.56	0.5	56.3	0.70	4.03	-0.17	0.72	-0.07	0.89	-0.24
Item6	0.53	0.51	53.1	0.56	1.97	-0.49	0.90	-0.02	1.04	0.27
Item7	0.84	0.37	84.4	0.47	1.78	-0.93	0.73	0.09	0.95	-0.01
Item8	0.5	0.51	50	0.39	1.15	-0.07	0.90	-0.39	0.95	-0.34
Item9	0.88	0.34	87.5	0.36	1.16	-1.09	0.87	0.07	1.09	0.34
Item10	0.38	0.49	37.5	0.54	2.23	-0.51	0.84	0.01	0.92	-0.30
Item11	0.47	0.51	46.9	0.58	2.25	-0.43	0.73	-0.17	0.92	-0.32
Item12	0.81	0.4	81.3	0.63	2.36	-1.11	1.03	0.35	1.11	0.42
Item13	0.97	0.18	96.9	0.36	5.93	-0.52	0.01	-3.04	0.10	-1.02
Item14	0.59	0.5	59.4	0.60	2.12	-0.35	0.73	-0.23	0.92	-0.30
Item15	0.56	0.5	56.3	0.61	2.54	-1.09	1.03	0.35	0.95	-0.13
Item16	0.63	0.49	62.5	0.47	1.40	-0.29	0.82	-0.44	0.95	-0.23
Item17	0.75	0.44	75	0.55	1.72	0.20	0.91	0.11	1.01	0.13
Item18	0.53	0.51	53.1	0.68	4.12	-0.99	1.71	0.97	0.92	-0.16
Item19	0.81	0.4	81.3	0.60	2.32	0.10	1.14	0.46	1.03	0.20
Item20	0.63	0.49	62.5	0.47	1.31	-1.08	0.92	-0.17	0.96	-0.16
Item21	0.66	0.48	65.6	0.65	2.87	-2.07	0.76	0.06	1.12	0.53
Item22	0.75	0.44	75	0.39	1.28	-0.29	0.91	-0.04	0.98	-0.00
Item23	0.69	0.47	68.8	0.68	2.96	-0.17	0.71	-0.04	0.94	-0.12
Item24	0.63	0.49	62.5	0.66	2.75	-0.49	0.55	-0.15	0.88	-0.41
Item25	0.84	0.37	84.4	0.65	3.68	-0.93	0.66	0.19	0.94	0.02
Item26	0.59	0.5	59.4	0.56	2.13	-0.07	1.03	0.26	0.98	-0.04
Item27	0.44	0.5	43.8	0.57	2.11	-1.09	1.02	0.25	1.03	0.24
Item28	0.78	0.42	78.1	0.58	2.07	-0.51	0.69	0.01	1.08	0.36

Model fit indices (Table 4) indicated an overall good fit to the IRT model, although some items presented potential misfit. For instance, Item3 showed major outfit (2.76) and infit (1.35) statistics, revealing inconsistent responses across ability levels, while Item13 exhibited extreme values due to its high facility rate (96.9%) despite a high discrimination estimate ($a = 5.60$) and

high item information within lower abilities. Nonetheless, most items showed outfit and infit statistics within acceptable bounds.

Figure 4

Test Characteristic Curve of the Tests Altogether



The Test Characteristic Curve (Figure 4) illustrated a smooth increase in the expected score on the whole test for participants at different levels within the expected range. This curve confirmed that the IRT model captured the relationship between ability and item performance consistently across the tests. The curve's slope was steepest at the mid-level of ability, reflecting the concentration of highly discriminating items in this range, and flattened at the extremes, further illustrating the limited precision in distinguishing very low- or very high-ability. The Test Information Function (Figure 5) and the Standard Error of Measurement (Figure 6) showed the same result with decreasing precision at the extremes. This limitation is a direct consequence of the scarcity of items with high difficulty or very low facility values.

Figure 5

Test Information Function for the Three Tests Altogether

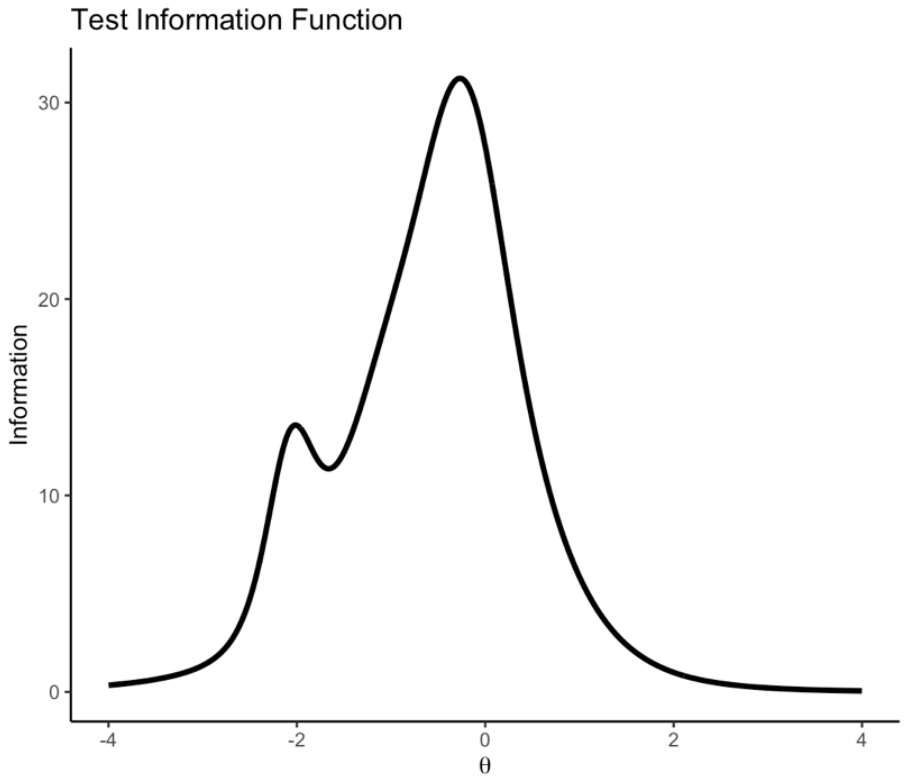


Figure 6

Standard Error of Measurement across Ability for the Three Tests Altogether

These findings align with cloze testing research by showing that discrimination tends to be strongest near the targeted difficulty, with limited precision at proficiency extremes (Oller & Conrad, 1971; Tremblay, 2011), and that text characteristics (e.g., cohesion, cultural familiarity) can influence item functioning (Trace, 2023). In line with the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), the results support the use of the test for group-level interpretations. The results hint that it might be possible to assess a broader range of proficiency using poetry-based cloze tests.

Logistic Regression Models on PFL Learner's Results

Apart from the psychometric models for test and item parameter estimates, we also applied a logistic regression model (binomial) to examine the influence of several sociolinguistic predictors on the probability of correct responses in the cloze tests (see Model 1 in Table 5). The predictors included country of residence, native language group, current class level (*A1*, *A2*, *B1*, *B2*), and self-assessed proficiency (*beginner*, *beginner-intermediate*, *intermediate*, *intermediate-advanced*). Country of residence was divided into two categories to optimize analysis: *abroad* and *Portugal*. The data for native language was divided into four groups: *Slavic languages* (Croatian, Polish, Russian, Serbian, Slovenian); *East Asian languages* (Mandarin, Japanese); *Romance/Germanic languages* (French, Spanish; German, respectively); *Western/Central Asian languages* (Persian, Turkish, Mongolian).

Table 5

Results from Logistic Regression Models Predicting the Probability of Correct Responses in the Three Cloze Tests Altogether

Variable	Estimate (β)	Std. Error	z value	p-value	Odds ratio
MODEL 1 (sociolinguistic variables)					
(Intercept)	-3.8219	1.7634	-2.167	0.0302	0.022
EP years	-0.7314	0.4826	-1.516	0.1296	0.481
weekly hours	0.1814	0.5036	0.360	0.7187	1.199
country of residence (Portugal)	1.9886	1.1193	1.777	0.0756	7.303
native language (Romance or Germanic)	0.9185	1.7173	0.535	0.5928	2.505
native language (Slavic)	3.5941	1.8269	1.967	0.0492	36.375
native language (Western and Central Asia)	1.4970	1.1953	1.252	0.2104	4.470
class level A2	-0.2547	0.9940	-0.256	0.7978	0.775
class level B1	-0.7935	1.8119	-0.438	0.6614	0.452
class level B2	-1.4156	1.7758	-0.797	0.4254	0.243
self-assessment (beginner-intermediate)	2.0953	0.9698	2.161	0.0307	8.126
self-assessment (intermediate)	3.7756	1.8383	2.054	0.0400	43.574
self-assessment (intermediate-advanced)	4.2353	1.7309	2.447	0.0144	68.978
MODEL 2 (structure and feedback variables)					
(Intercept)	1.5472	0.5784	2.675	0.0075	4.699
structure (G2_PRO_R_P_I)	-0.0439	0.6161	-0.071	0.9432	0.957
structure (G3_PREP_CONTR)	-0.6438	0.6413	-1.004	0.3154	0.526
structure (G4_DET_N_MODIF)	0.4586	0.6492	0.706	0.4800	1.582
structure (G5_PRO_CLIT_POLA)	-0.1920	0.7061	-0.272	0.7857	0.825
feedback:test_1	-0.7509	0.3435	-2.186	0.0288	
feedback:test_2	-0.6846	0.3866	-1.771	0.0766	
feedback:test_3	-0.9149	0.3722	-2.458	0.0140	

Reference levels were set as *abroad* for country residence, *East Asian languages* for native language, *A1* for class level, and *beginner* for self-assessment. Random intercepts were

included for *participant* and *item*. All numeric continuous variables were centered, which ensures that the coefficients are comparable, reduces multicollinearity, improves interpretability of main effects, and reduced model convergence issues. This model only includes 28 participants (out of 32), due to the exclusion of participant entries with missing values for the respective variables in analysis. Note that this issue occurs solely for the sociolinguistic variables, since providing that information was not made a requirement to participate in the completion of the cloze tests.

Self-assessment emerged as a significant predictor, since participants who rated themselves as *intermediate* or *intermediate-advanced* had significantly higher odds of answering correctly than those who rated themselves as *beginner*. The effect was strongest for the intermediate-advanced group ($\beta = 4.24$, $p = 0.0144$). Native language also showed an effect. Participants from *Slavic language* backgrounds performed significantly better ($\beta = 3.59$, $p = 0.049$), compared to the reference group. No significant differences were found for *Romance/Germanic* or *West/Central Asian* languages. In the case of country of residence, participants living in *Portugal* showed marginal effect ($\beta = 1.99$, $p = 0.076$), which entails that immersion might improve performance, although this was not statistically conclusive.

The random intercept for participants (SD = 1.03) shows substantial differences in individuals' baseline likelihood of answering correctly, even after controlling for feedback and test type. This variability translates into wide probability ranges, in some cases, differences of roughly a quarter to a third of the scale between lower- and higher-performing participants. The random intercept for items (SD = 0.97) reveals that certain items are consistently easier or harder than others, with variability similar in magnitude to that seen across participants. These results show that both participant ability and item difficulty are major sources of variation in scores, thus including these random effects was essential to avoid underestimating uncertainty or inflating the apparent significance of fixed effects.

The variable concerning the number of years of learning (years of exposure to EP) did not emerge as a statistically significant predictor of performance in the cloze tests ($\beta = -0.73$, $p = 0.13$), contrarily to our expectation. Interestingly, the negative estimate reveals a trend in which longer exposure to EP is associated with lower accuracy, although this relationship is not strong enough to be conclusive. One possible interpretation is that years of exposure may be acting as a proxy for other underlying learner characteristics not accounted for in this study, or inefficient item design.

The results showed significant main effects for multiple predictors:
Country of residence: Participants residing in Portugal had a notably higher probability of correct responses (13.9%) compared to the baseline group (2.2%), reflecting a positive effect of immersion or environment on language performance.

Native language: Speakers of Slavic languages showed the highest increase in probability (+41.2%) compared to the baseline group, followed by West/Central Asian (+7.1%) and Romance/Germanic (+3.1%) speakers.

Class level: Surprisingly, higher class levels (A2, B1, B2) showed a slight decrease in probability compared to the baseline A1 group, while the absolute probabilities remained low (below 2%). This counterintuitive result may reflect sample characteristics or measurement issues.

Self-assessment: Participants rating themselves as intermediate-advanced showing a large increase in probability (60.2%) relative to beginners (2.2%). This suggests that learners' self-perceptions align well with actual performance, underscoring the validity evidence provided by self-assessment as a proxy for proficiency in this context.

Thus, participants classified in the baseline categories – residing outside Portugal, with the reference native language group, at class level A1, and self-assessed as beginners – had a very low predicted probability of a correct response (2.17%). This low baseline probability sets the context for evaluating the effect of other factors. No significant interactions between predictors were found in the current model. In addition, the model showed acceptable fit statistics (AIC = 759.9, BIC = 829.9). Model assumptions were checked and met, ensuring accurate interpretation of results and explaining variability in performance. The sample size and diversity limit generalizability.

A second logistic regression model (see Model 2 in Table 5) investigated how different linguistic structure groups and feedback on test difficulty (0 = "easy" vs. 1 = "difficult") via the interaction between feedback and the different tests (test 1, test 2, and test 3), influenced the probability of correct responses. The baseline structures (G1) under baseline feedback (0) served as the reference point for comparisons. The model included fixed effects for *structure* type and interaction term *feedback:test*, along with random intercepts for *participant* and *item*. Each coefficient is already the within-test feedback effect. No participants were excluded from this model (all 32 were included).

Test version showed a strong effect, as participants performed significantly worse in test 1 and test 3, suggesting variation in test difficulty or in how *feedback* was incorporated (*test 1: β = -0.75, p = 0.029; test 3: β = -0.91, p = 0.014*). Test 2 showed a marginal effect (β = -0.68, p = 0.077), pointing in the same direction. In all three tests, feedback "difficult" was associated with a lower log-odds of correct responses relative to the same test with feedback "easy". Structure type did not affect performance, considering that none of the specific linguistic structure groups showed a significant effect. The directions of the effects vary (some positive, some negative), but due to large standard errors, these differences could easily be due to random variation rather than real performance differences between structure groups. This indicates that no individual group of structures was inherently more difficult, once *participant* and *item* variance were accounted for. The variance in participants' random effects was higher in this model (SD = 1.69) than in the sociolinguistic model (SD = 1.03), suggesting more unexplained participant-level variability in this model.

The different structure groups showed varied main effects on the probability of correct responses:

(i) The baseline structure group (G1) had a high probability of correct responses (82.8%) under baseline feedback (0), serving as the standard for comparison.
(ii) Some structures, such as in G4, exhibited higher probabilities (88.8%) than baseline (G1), suggesting they may be easier or more familiar to test takers.
(iii) Conversely, G3 showed a substantial decrease in probability (71.66%), representing higher difficulty.

The analysis of structure group complexity did not reveal significant performance differences between the five structure groups. This partially contrasts with previous studies (Flores et al., 2022; Fiéis et al., 2023; Rinke et al., 2024), which identified clitic pronouns, contracted prepositions, and relative pronouns as particularly complex and dependent on extensive exposure to written registers. In our data, for example, structures such as contracted prepositions (G3) showed a tendency toward lower probabilities of correct responses compared to baseline, aligning with the notion of higher processing demands, but the effect was not statistically robust. Likewise, clitic pronouns (G5) did not emerge as systematically more difficult, which may reflect limited statistical power or the influence of task-specific features such as context.

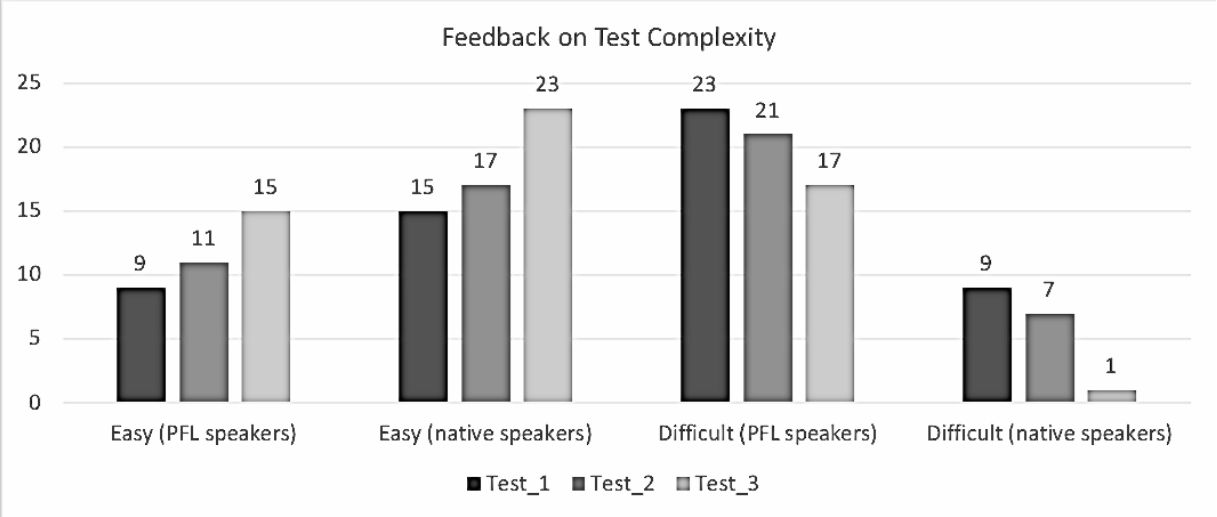
Perception of Test Difficulty

Regarding the perception of the difficulty of each cloze test (see Figure 7), the pretesting results from native speakers show that 15 speakers found Test 1 easy, while 9 found it difficult. For Test 2, 17 rated it as easy, and 7 as difficult, whereas Test 3, was perceived as the easiest, with 23 participants finding it easy and only 1 finding it difficult. The higher difficulty perception of Test 1 aligns with its highest error rate (55.71%) and the presence of higher complexity items (e.g., [*por*], [*que*], [*com*]). The higher number of items may have also influenced this perception.

Similarly, PFL speakers followed the same pattern (see Figure 7). In Test 1, 9 (28%) found it easy, while 23 (72%) considered it difficult. For Test 2, 11 (34%) rated it as easy, while 21 (66%) found it difficult. Test 3 was perceived as the least difficult, although 15 (47%) found it easy and 17 (53%) found it difficult. The greater difficulty reported in Test 1 and 2 is likely due to their more complex items. Despite these perceptions, score variability across tests was minimal.

Figure 7

Participants' Perception of Difficulty for each Cloze Test



Interpretations and Limitations of Key Findings

Can poetry-based cloze tests be used as a viable language assessment tool for EP as a foreign language?

The findings demonstrate that a poetry-based cloze test can achieve strong psychometric performance, high internal consistency and robust item discrimination within the lower-to-intermediate proficiency range. Contrary to expectations that poetry would render the task excessively difficult, the items clustered around moderate difficulty, showing that participants were able to mobilize contextual, semantic, lexical, and syntactic cues effectively to fill in the gaps. There is no strong impediment or determinant factor that compromises the functionality and practicability of poetry within the present test construct. These contributions entail that literary texts can broaden the construct coverage of cloze procedures, but they also demand psychometric rigor with interpretive openness.

Using poetry adds distinct cognitive demands, such as condensed syntax, ambiguity, and reliance on stylistic cues such as rhythm, rhyme, and metaphor. These features require readers to engage in higher-order inferencing, relying not only on linguistic competence but also literary awareness. This can strengthen construct validity by reflecting authentic language use, but it also creates challenges. Multiple possible responses may weaken control, and the interpretive openness of poetry can blur the line between assessing language proficiency and assessing literary interpretation skills if not carefully controlled. Moreover, it raises the perennial question in language testing of whether cloze tests assess “reading comprehension,” “linguistic competence,” or a broader construct that includes cognitive sensitivity.

Can the target linguistic structure groups being assessed predict participants overall performance separately?

The target linguistic structure groups showed limited predictive power for performance – perhaps due to the low number of items or the established grouping of structures. Instead, outcomes were likely driven by item-design features that controlled difficulty and kept the

construct consistency across forms. Letter-count hints per item in Tests 1 and 2 moderated their complexity, despite their less prose-like nature, while Test 3, lacking such clues, relied on a more intuitive, prose-based assessment. Despite being predicted as easier, Test 3's item design balanced item complexity across tests. Thus, by manipulating the amount of information provided about the correct response of each item, we were able to control difficulty. This signifies that tailoring procedures for a particular proficiency level through task and response format design is possible (see also Brown et al., 2001), although with certain limitations and rules for different populations – in this case (young) adults. The inclusion of these "hints" likely increased facility but did not induce correct responses, rather only verified them. While poetry-based cloze tests should not replace standard prose-based assessments, they provide a complementary, creative, and engaging alternative. Finally, these assessments contribute to test fairness discussions (Kunnan, 2004) maintaining construct validity while reducing stress associated with traditional methods (Kanonidou & Papachristou, 2019).

The results of this study also suggest that the complexity of poetry texts for assessment purposes may have been underestimated**.** We expected poetry-based cloze items to increase the complexity of the tests due to text type and format. Contrary to this expectation, virtually no items functioned as genuinely difficult tasks. This indicates that, even within poetry, test takers were able to mobilize contextual, semantic, and stylistic cues to infer the missing words. The implication is that item selection could have been more ambitious, with the inclusion of items deliberately designed to push more advanced test takers alongside easier items to capture lower proficiency.

Is there a correlation between test-taker' scores and their sociolinguistic profile?
Self-assessment persists as one of the most consistent predictors of test performance – phenomenon documented in studies, while assuming variation depending on experience and learner profile (Ross, 1998). In fact, participants' perceptions inform task design and can contribute to test validity arguments (Cheng & DeLuca, 2011).

Some limitations of this study include uneven learner distribution and a lower-than-recommended item count. Sociolinguistic factors were considered, but cognitive influences warrant further study. Furthermore, due to missing data in sociolinguistic variables, we conducted two separate models: one with all participants' data to examine test- and item-related effects, and another using the reduced dataset to examine sociolinguistic factors' influences.

Conclusion

This study shows that poetry-based cloze tests are a viable, classroom-friendly option for assessing EP at beginner-to-intermediate levels. As one of the first studies to explore poetry in cloze tests for PFL assessment, this research lays a foundation for further investigations and offers practical insights for teachers, especially at a local level, for the assessment of language proficiency. Future work should build on these insights by exploring and diversifying item difficulty, exploring different poetic genres, and finding validity evidence in results of a larger sample, thereby redefining the role of poetry in language testing and advancing the validity evidence base for innovative assessment practices.

ORCID

https://orcid.org/0009-0000-2380-1466

https://orcid.org/0000-0001-6173-8605

Acknowledgments

First and foremost, we would like to express our sincere gratitude to Professor Doctor Cristina Flores for her constructive feedback and her valuable advice on methodological matters. We would like to extend our thanks to all the participants in this study for their time and input.

Funding

This work was supported by FCT - Fundação para a Ciência e Tecnologia, I.P. by project reference 2022.09752.BD and DOI identifier https://doi.org/10.54499/2022.09752.BD.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

References

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, *13*(2), 219–227. https://doi.org/10.2307/3586211

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732935

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Armon-Lotem, S. (2005). The acquisition of subordination: From preconjunctionals to later use. In D. Ravid & H. B.-Z. Shyldkrot (Eds.), *Perspectives on language and language development* (pp. 191–202). *Springer*. https://doi.org/10.1007/1-4020-7911-7_15

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *16*(1), 61–70. https://doi.org/10.2307/3586277

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Batalha, J. (2018). *Relações entre o conhecimento explícito da língua e a competência de leitura. Compreensão de dependências referenciais no ensino básico.* [Doctoral dissertation, Universidade NOVA de Lisboa]. http://hdl.handle.net/10362/43439

Brown, A. (1993). The role of test-taker feedback in the test development process: Test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing*, *10*(3), 277–301. https://doi.org/10.1177/026553229301000305

Brown, J. D., Yamashiro, A. D., & Ogane, E. (2001). The emperor's new cloze: Strategies for revising cloze tests. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development* (pp.143–161). University of Hawai'i Press.

Chapelle, C. A. & Abraham, R. G. (1990). Cloze method: What differences does it make? *Language Testing*, *7*(2), 121–146. https://doi.org/10.1177/026553229000700201

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.) (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.

Charneca Catalão, M. F. (2011). Os padrões de uso dos pronomes pessoais àtonos em português europeu. University of Évora. [Mater dissertation, Universidade de Évora]. http://hdl.handle.net/10174/15325

Cheng, L. (1997). How Does Washback Influence Teaching? Implications for Hong Kong. *Language and Education*, *11*(1), 38–54. https://doi.org/10.1080/09500789708666717

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment*, *16*(2), 104–122. https://doi.org/10.1080/10627197.2011.584042

Chung, E. S., & Ahn, S. (2019). Examining cloze tests as a measure of linguistic complexity in L2 writing. *Language Research, 55*(3), 627–649. https://doi.org/10.30961/lr.2019.55.3.627

Costa, J., Fiéis, A. & Lobo, M. (2015). Input variability and late acquisition: Clitic misplacement in European Portuguese. *Lingua*, 161, 10–26. https://doi.org/10.1016/j.lingua.2014.05.009

Costa, A. L. (2006). Complexidade estrutural de conectores concessivos. In *Textos selecionados do XXII Encontro da Associação Portuguesa de* Linguística, 207–302. Associação Portuguesa de Linguística.

Costa, J. & Lobo, M. (2009). Clitic omission in the acquisition of European Portuguese: Data from comprehension. In A. Pires & J. Rothman (Eds.), *Minimalist inquiries into child and adult language acquisition: Case studies across Portuguese*, 63–84. Mouton de Gruyter. https://doi.org/0.1515/9783110215359.1.63

Costa, J., Lobo, M. & Silva, C. (2012). Which category replaces an omitted clitic? The case of European Portuguese. In P. Larrañaga & P. Guijarro- Fuentes (Eds.), *Pronouns and clitics in early language*, (Vol. 108, pp. 105–130). De Gruyter, Inc. https://doi.org/10.1515/9783110238815.105

Costa, J., Lobo, M. & Silva, C. (2011). Subject–object asymmetries in the acquisition of Portuguese relative clauses: Adults vs. children. *Lingua*, 121(6). 1083–1100. https://doi.org/10.1016/j.lingua.2011.02.001

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume.* Council of Europe Publishing.

Dabarera, C., Renandya, W. A., & Zhang, L. J. (2014). The impact of metacognitive scaffolding and monitoring on reading comprehension. *System*, *42*, 462–473. https://doi.org/10.1016/j.system.2013.12.020

Direção de Serviços de Língua e Cultura (2017). *Referencial Camões PLE*. Camões, Instituto da Cooperação e da Língua I.P.

Effatpanah, F., Baghaei, P., Tabatabaee-Yazdi, M., & Babaii, E. (2024). A new scoring method for item response theory analysis of C-Tests. *Language Testing*, *42*(2), 167–192. https://doi.org/10.1177/02655322241265350

Ehret, K. & Szmrecsanyi, B. (2016). An information-theoretic approach to assess linguistic complexity. In R. Baechler & G. Seiler (Ed.), *Complexity, isolation, and variation*, 71–94. De Gruyter. https://doi.org/10.1515/9783110348965-004

Elatia, S. (2011). Choosing language competence descriptors for language assessment: Validity and fairness issues. *Synergies Europe*, *6*, 165–175. https://gerflint.fr/Base/Europe6/samira.pdf

Espanca, F. (2012). *Livro de mágoas* (C. P. Alonso & F. M. da Silva, Eds.). Estampa.

Fabb, N. (2015). *What is poetry? Language and memory in the poems of the world*. Cambridge University Press. https://doi.org/10.1017/CBO9780511736575

Farrah, M., & Al-Bakri, R. (2022). The effectiveness of using poetry in developing English vocabulary, pronunciation and motivation of EFL Palestinian students. *Language Teaching*, *2*(1). https://doi.org/10.30560/lt.v2n1p1

Férez Mora, P. A., & Coyle, Y. (2020). Poetry for EFL: Exploring change in undergraduate students' perceptions. *Porta Linguarum, 33*, 231–247. http://hdl.handle.net/10481/62829

Fiéis, A., Madeira, A., & Teixeira, J. (2023). Colocação de clíticos em PE L2. *Revista da Associação Portuguesa de Linguística*, (10), 115–137. https://doi.org/10.26334/2183-9077/rapln10ano2023a7

Flores, C, Rinke, E. & Sopata, A. (2020). Acquiring the distribution of null and overt direct objects in European Portuguese. *Journal of Portuguese Linguistics*, *19*(5), 1–20. https://doi.org/10.5334/jpl.239

Flores, C., Gonçalves, M. L., Rinke, E., & Torregrossa, J. (2022). Perspetivas múltiplas sobre a competência bilingue de crianças lusodescendentes residentes na Suíça: A investigação linguística em diálogo com a didática. *Revista Portuguesa de Educação, 35*, 102–131. https://doi.org/10.21814/rpe.24205

Flores, C., Setas, C., Sousa, I. & Antunes, J. (2021). Ein Cloze Test zur Spachstandserhebung portugiesischer L2 Lerner des Deutschen. *POLISSEMA – Revista de Letras do ISCAP*, 21. http://hdl.handle.net/10400.22/20763

Fotos, S. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning*, *41*(3), 313–336. https://doi.org/10.1111/j.1467-1770.1991.tb00609.x

Gardihewa, P. N. (2022). Use of poetry in the English as a second language classroom: A study of second year undergraduates in Sabaragamuwa University of Sri Lanka. *Vidyodaya Journal of Humanities and Social Sciences, 7*(1), 85-104. http://doi.org/10.31357/fhss/vjhss.v07i01.05

Giraldo, F. D. (2020). Validity and classroom language testing: A practical approach. *Colombian Applied Linguistics Journal*, *22*(2), 194–206. https://doi.org/10.14483/22487085.15998

Green, S. B., & Yang, Y. (2009). Reliability of summed item scores using structural equation modeling: An alternative to coefficient alpha. *Psychometrika*, *74*, 155–167. https://doi.org/10.1007/s11336-008-9099-3

Hartley, J., & Trueman, M. (1986). The effects of the typographic layout of cloze-type tests on reading comprehension scores. *Journal of Research in Reading*, *9*(2), 116–124. https://doi.org/10.1111/j.1467-9817.1986.tb00118.x

Hossain, K. I. (2024). Literature-based language learning: Challenges, and opportunities for English learners. *Ampersand*. *13*. https://doi.org/10.1016/j.amper.2024.100201

Iskandarova, G. (2024). Current issues in language assessment and language assessment research and its implication. *Baltic Journal of Legal and Social Sciences*, *3*, 228–231. https://doi.org/10.30525/2592-8813-2024-3-24

Kalyoncu, M., & Memiş, M. (2025). Cloze tests in measuring reading comprehension levels. *Education and Science*, *50*, 69–92. https://doi.org/10.15390/EB.2025.14079

Kanonidou, E., & Papachristou, V. (2019). The use of songs, lyrics and poetry in EFL teaching and in SLA: Students' and teachers' views. In C. Can, P. Patsala, & Z. Tatsioka (Eds.), *Language in focus: Contemporary means and methods in ELT and applied linguistics* (pp. 331–354). LIF – Language in Focus Publications.

Kellem, H. (2009). The Formeaning response approach: Poetry in the EFL classroom. *English Teaching Forum*, *47*(4), 12–17.

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

Le, V.-N., & Klein, S. P. (2002). Technical criteria for evaluating tests. In S. P. Klein, L. S. Hamilton, & B. M. Stecher (Eds.), *Making sense of test-based accountability in education*, (1st ed., pp. 51–78). RAND Corporation. http://www.jstor.org/stable/10.7249/mr1554edu.10

Lee, L., & Lin, S. C. (2015). The impact of music activities on foreign language, English learning for young children. *Journal of the European Teacher Education* Network, *10*, 13–23.

Lobato de Faria, R. (1996). *Poemas escolhidos e dispersos*. Editorial Caminho.

Lousada, M., Mendes, A. P., Valente, A. R., & Hall, A. (2012). Standardization of a phonetic-phonological test for European-Portuguese children. *Folia Phoniatrica et Logopaedica*, *64*(3), 151–156. https://doi.org/10.1159/000264712

Malheiros Teodoro, G. (2020). A aquisição das preposições de, para e com por falantes do português do Brasil. [Bachelor's thesis, Universidade de Brasília]. https://bdm.unb.br/handle/10483/26283

Martínez, C. (2008). La enseñanza de los contenidos culturales a través de la poesía. [Master's thesis, Universidad Complutense de Madrid]. Repositorio de la Biblioteca Virtual, Ministerio de Educación y Formación Profesional. https://www.educacionfpydeportes.gob.es/mc/redele/biblioteca-virtual/numerosanteriores/2008/memoriamaster/2-semestre/martinez-c.html

McNamara, T. (2001). Language assessment as social practice: challenges for research. *Language Testing*, *18*(4), 333–349. https://doi.org/10.1177/026553220101800402

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (Eds.), *Competence and performance in language learning*, 35–53. Cambridge University Press.

Mermelstein, A. D. (2022). Benefits of poetry: An argument for making poetry a required course for EFL literature majors. *Curriculum Matters*, *18*, 27–45. https://doi.org/10.18296/cm.0058

Mésseder, J. P. (2003). *O G é um Gato Enroscado*. Editorial Caminho.

Mittal, R. (2014). Teaching English through Poetry: A Powerful Medium for Learning Second Language. *IOSR Journal of Humanities and Social Science*, *19*(5), 21–23. Retrieved from https://www.iosrjournals.org/iosr-jhss/papers/Vol19-issue5/Version-3/D019532123.pdf

Nation, P. (2002). Best Practice in Vocabulary Teaching and Learning. In J. C. Richards & W. A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice*, 267–272. Cambridge University Press.

Oliveira, B. (2018). O texto em verso no desenvolvimento da competência lexical em PLE. [Master's thesis, Faculdade de Letras da Universidade do Porto]. https://hdl.handle.net/10216/116896

Oller, J. W. Jr., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning*, *21*, 183–195. https://doi.org/10.1111/j.1467-1770.1971.tb00057.x

Oller, J. W. Jr., & Jonz, J. (1994). A critical appraisal of related cloze research. In J. W. Oller, Jr., & Jon Jonz (Eds.), *Cloze and coherence*, 371–408. Bucknell University Press. https://doi.org/10.13140/RG.2.1.3179.1844

Oller, J. W. Jr. (1979). *Language tests at school: A pragmatic approach*. Longman.

Park, C.-Y. (2011). Making cloze tests more valid. *Journal of the Korea Academia Industrial Cooperation Society*, *12*(2), 640–645. https://doi.org/10.5762/KAIS.2011.12.2.640

Peskin, J. (2007). The genre of poetry: Secondary school students' conventional expectations and interpretive operations. *English in Education*, *41*(3), 20–36. https://doi.org/10.1111/j.1754-8845.2007.tb01162.x

Plaza, C. F. (2009). Poesía en la clase de ELE: Propuestas didácticas. In *V Encuentro Brasileño de Profesores de Español*, (9). Retrieved from https://marcoele.com/descargas/enbrape/ferrer_poesia.pdf

Purpura, J. E. (1999). *Learner strategy use and performance on language tests*. Cambridge University Press. Vol. 8.

Rahimi, S. (2014). Cloze test: Form a testing instrument toward a tool in SLA, new applications. *Journal of Academic and Applied Studies*, *4*(6), 26–39. Retrieved from https://www.academia.edu/8522169

Reazul, M. (2022). Literature in EFL/ESL classroom: Integrating conventional poetry as authentic material. *International Journal of Language and Literary Studies*, *4*(3), 312–328. http://doi.org/10.36892/ijlls.v4i3.1052

Ribeiro, J. M. (2007). O valor pedagógico da poesia. *Revista Portuguesa De Pedagogia*, *41*(2), 51–81. https://doi.org/10.14195/1647-8614_41-2_3

Rinke, E., Flores, C. & Torregrossa, J. (2024). How different types of complexity can account for difficult structures in bilingual and monolingual language acquisition. In M. Polinsky & M. Putnam (eds.), *Formal approaches to complexity in heritage languages*, 43–71. Language Science Press. https://doi.org/10.5281/zenodo.12090434

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), 242–258. https://doi.org/10.1111/j.1473-4192.2006.00117.x

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*(1), 1–20. https://doi.org/10.1177/026553229801500101

Sánchez, R. L. (2017). La enseñanza de la poesia en el aula de español como lengua extranjera. *Editorial Síntesis*. https://www.sintesis.com/libro/la-ensenanza-de-la-poesia-en-el-aula-de-espanol-como-lengua-extranjera

Santos, A. L. (2017). Alguns aspetos da aquisição de orações subordinadas completivas. In M. J. Freitas & A. L. Santos (Eds.), A aquisição de língua materna e não materna: Questões gerais e dados do português, 249–273. *Language Science Press*. https://doi.org/10.5281/zenodo.889436

Schander, C., Balma, B. M., & Massa, A. A. (2013). The joy of art in the EFL classroom. In *European Multidisciplinary Forum 2014:* Proceedings, Vol. 2, 409–414. European Scientific Institute. https://eujournal.org/files/journals/1/books/emf.vol.2.pdf

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *531*(1), 109–120. https://doi.org/10.1017/S0261444819000326

Trace, J., Brown, J. D., Janssen, G., & Kozhevnikova, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing*, *34*(2), 151–174. https://doi.org/10.1177/0265532215623581

Trace, J. (2023). The influence of passage cohesion on cloze test item difficulty. *Language Teaching Research Quarterly*, *37*, 161–178. https://doi.org/10.32038/ltrq.2023.37.08

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, *33*(3), 339–372. https://doi.org/10.1017/S0272263111000015

Tremblay, A., & Garrison, M. D. (2010). Cloze tests: A tool for proficiency assessment in research on L2 French. In Matthew T. Prior et al. (Eds.), Selected Proceedings of the 2008 *Second Language Research Forum* (pp. 73–88). Somerville, Cascadilla Proceedings Project.

VanPatten, B. (2015). Input Processing in Adult Second Language Acquisition. In B. VanPatten, & J. Williams (Eds.), *Theories in second language acquisition: An introduction*, 113–134. Routledge. https://doi.org/10.4324/9780429503986-6

Vardell, S. M., Hadaway, N. L., & Young, T. A. (2011). Matching books and readers: Selecting literature for English learners. *The Reading Teacher*, *59*(8), 734–741. https://doi.org/10.1598/RT.59.8.1

Vasconcelos, M. (1995). Relative clause acquisition and experimental re-search: A study with Portuguese children. In Isabel Hub Faria & Maria João Freitas (eds.), *Studies on the acquisition of Portuguese*, 115–128. APL.

Wakid, M., Sofyan, H., Widowati, A., & Zaida Ilma, A. (2024). Learning-oriented assessment: A systematic literature network analysis. *Cogent Education*, *11*(1). https://doi.org/10.1080/2331186X.2024.2366075

Wall, D. (2005). *The impact of high-stakes examinations on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge University Press. https://doi.org/10.1016/S0346-251X(00)00035-X

Watanabe, Y., & Koyama, D. (2008). A meta-analysis of second language cloze testing research. *Second Language Studies*, *26*(2), 103–133. http://hdl.handle.net/10125/40694

Weaven, M., & Clark, T. (2013). "I guess it scares us": Teachers discuss the teaching of poetry in senior secondary English. *English in Education*, *47*(3), 197–212. https://doi.org/10.1111/eie.12016

Wilkins, D. A. (1972). *Linguistics in language teaching*. Edward Arnold.

Wright, R. R., Coryell, J. E., Martinez, M., Harmon, J., Henkin, R., & Keehn, S. (2010). Rhyme, response, and reflection: An investigation of the possibilities for critical transformative learning through adult poetry reading. *Journal of Transformative Education*, *8*(2), 103–123. https://doi.org/10.1177/1541344611406737

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, *20*(3), 267–293. https://doi.org/10.1191/0265532203lt257oa

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*(1), 21–29. https://doi.org/10.22237/jmasm/1177992180

Appendices

Appendix 1

Table A - Items in Cloze Test 1 (Poem: Um livro)

Item	Passage	Gap	Linguistic structure	Structure ID
1	Levou-me um livro em viagem não sei por onde é que andei Corri o Alasca, o deserto andei ___ o sultão no Brunei? P'ra falar verdade, não sei	[com]	Simple preposition	G3_PREP_CONTR
2	Com um livro cruzei o mar, não sei com ___ naveguei. Com marinheiros, corsários, tremendo de febres e medo? P'ra falar verdade não sei.	[quem]	Relative pronoun invariable	G2_PRO_R_P_I
3	Um livro levou-me p'ra longe não sei por onde é que and___.	and[ei]	Verb conjugation: 1st person, singular, past tense, indicative	G1_V_INFL
4	Por cidades devasta___ no meio da fome e da guerra? P'ra falar verdade não sei.	devasta[das]	Adjective inflection: feminine, plural	G4_DET_N_MODIF
5	Um livro levou-me com ele até ___ coração de alguém	[ao]	Preposition contraction with a definite article: 'a' + 'o', masculine, singular	G3_PREP_CONTR
6	E aí me enamorei – de ___ olhos	[uns]	Indefinite article: masculine, plural	G4_DET_N_MODIF
7	ou de uns cabel___? P'ra falar verdade não sei.	cabel[os]	Noun inflection: masculine, plural	G4_DET_N_MODIF
8	Um livro num passe de mágica tocou-me com o ___ feitiço:	[seu]	Possessive pronoun: masculine, singular	G2_PRO_R_P_I
9	Deu-me a paz e deu-me a guerra, mostrou-___ as faces do homem – porque um livro é tudo isso.	mostrou-[me]	Clitic pronoun: dative mode, 1st person, singular	G5_PRO_CLIT_POLA
10	Levou-me um livro com ele ___ mundo a passear	[pelo]	Preposition contraction with a definite article: 'por' + 'o', masculine, singular	G3_PREP_CONTR
11	___ me perdi nem me achei – porque um livro é afinal…	[Não]	Adverb of negation	G5_PRO_CLIT_POLA

12	um pouco ___ vida, bem sei.	[da]	Preposition contraction with a definite article: 'de' + 'a', feminine, singular	G3_PREP_CONTR

Table B - Items in Cloze Test 2 (Poem: Perdi os Meus Fantásticos Castelos)

Item	Passage	Gap	Linguistic structure	Structure ID
13	Perdi meus fantástic___ castelos	fantástic[os]	Adjective inflection: masculine, plural	G4_DET_N_MODIF
14	Como névoa distante ___ se esfuma...	[que]	Relative pronoun: subordinate adverbial clause of manner	G2_PRO_R_P_I
15	Quis vencer, quis lutar, quis defendê-___: Quebrei as minhas lanças uma a uma!	defendê-[los]	Clitic pronoun allomorph: accusative mode, masculine, plural	G5_PRO_CLIT_POLA
16	Perdi minhas galeras entre os gelos Que se afundar___ sobre um mar de bruma...	afundar[am]	Verb 3rd person, plural, past perfect, indicative	G1_V_INFL
17	- Tantos escolhos! ___ podia vê-los? – Deitei-me ao mar e não salvei nenhuma!	[Quem]	Interrogative pronoun	G2_PRO_R_P_I
18	Perdi a minha taça, o meu anel, A minha cota de aço, o meu corcel, Perdi meu elmo ___ ouro e pedrarias...	[de]	Simple preposition	G3_PREP_CONTR
19	Sobem-me aos lábi___ súplicas estranhas... Sobre o meu coração pesam montanhas...	lábi[os]	Noun inflection: masculine, plural	G4_DET_N_MODIF
20	Olho assombrada as ___ mãos vazias...	[minhas]	Possessive pronoun: feminine, plural	G2_PRO_R_P_I

Table C - Items in Cloze Test 3 (Poem: Imaginação)

Item	Passage	Gap	Linguistic structure	Structure ID
21	A imaginação é magia e é arte ___ nos faz inventar,	[que]	Relative pronoun: subordinate restrictive adjective clause	G2_PRO_R_P_I
22	sonhar e via___.	via[jar]	Verb inflection: infinitive	G1_V_INFL
23	Com imaginação podemos ir a Marte ou ao centro da Terra, ou ___ fundo do mar.	[ao]	Preposition contraction with a definite article: 'a' + 'o', masculine, singular	G3_PREP_CONTR
24	Com imaginação ___ estamos sozinhos.	[nunca]	Adverb of negation	G5_PRO_CLIT_POLA
25	A imaginação é um voo, um lugar ___ temos amigos,	[onde]	Relative pronoun: subordinate adverbial clause of place	G2_PRO_R_P_I
26	onde há outros caminhos nos quais, sem te mexer___, podes ir passear.	mexer[es]	Verb conjugation: 2nd person, singular, inflected infinitive	G1_V_INFL
27	Inventa uma cantiga, ___ poema, um desenho um arco-íris, um rio por entre malmequeres esse lugar é teu, sem limite ou tamanho.	[um]	Indefinite article: masculine, singular	G4_DET_N_MODIF
28	A esse teu lugar, só vai quem tu quiser___.	quiser[es]	Verb conjugation: 2nd person, singular, inflected infinitive	G1_V_INFL