Jose Fabián Elizondo-González[1]
Universidad de Costa Rica, San Jose
Peyman Jahanbin[2]
University of Kansas, Lawrence

Original scientific paper
UDC: 371.26:811.111
http://doi.org/10.5937/IstrPed2602146F

==========

A LINEAR LOGISTIC TEST MODEL (LLTM) APPLICATION IN FOREIGN LANGUAGE TESTING

Abstract: This study applies the Linear Logistic Test Model (LLTM) to the reading comprehension subtest of an English certification exam developed by the Foreign Language Assessment Program (PELEx, for its acronym in Spanish). A Q-matrix operationalized cognitive and linguistic predictors, such as paraphrasing and inferential reasoning, to explain item difficulty. Delta squared results showed that the Q-matrix accounted for 77% of the variance in item difficulty, with "Inferences" and "Subtask Complexity" being key contributors. While overlaps in item difficulty coefficients reflected the nested nature of the Common European Framework of Reference for languages (CEFR) levels, this progression aligns with the framework's principles. The findings show that LLTM can make item difficulty more interpretable in reading assessment while also helping identify which item features merit further refinement in future test development.

Keywords: fairness, IRT, language testing, LLTM, Q-matrix.

Introduction

Reading comprehension is among the cognitively demanding constructs to evaluate in high-stakes language assessments. It requires test-takers to decode linguistic structures, integrate textual meaning, and draw inferences across multiple levels of discourse. These processes become even more complex in second or foreign-language contexts, where readers must also overcome limited lexical access, weaker syntactic processing, and unfamiliar text structures (Grabe, 2008). As a result, understanding what makes a reading item difficult is essential not only for accurate measurement but also for promoting fairness, transparency, and instructional relevance.

Given the multidimensional nature of reading comprehension, particularly in L2 settings (Grabe, 2008) and the need for valid difficulty estimation, robust measurement models are needed. Item Response Theory (IRT) has become the dominant psychometric framework in this regard, offering methodological advantages such as item-level comparability and ability scaling (Embretson & Reise, 2000). Among IRT models, the Rasch model (Rasch, 1960) is especially popular in language testing because of its simplicity and desirable properties, such as specific objectivity and invariant item calibration.

However, despite these strengths, traditional IRT models - including Rasch - treat test items as statistical entities, inferring difficulty solely from response patterns without regard to the underlying cognitive or linguistic features that contribute to item complexity. This "black-box" treatment limits interpretability, especially for stakeholders such as teachers, curriculum designers,

[1] josefabian.elizondo@ucr.ac.cr; https://orcid.org/0000-0003-4819-0213
[2] peyman@ku.edu; https://orcid.org/0000-0002-5692-3379

and examinees, who seek to understand what makes an item more or less difficult. For test developers, this lack of specificity poses challenges when reporting the elements that influence difficulty or aligning items with curriculum standards.

To bridge this interpretive gap, the Linear Logistic Test Model (LLTM; Fischer, 1973) - a cognitively informed extension of the Rasch model - offers a promising alternative. LLTM improves upon traditional IRT models by decomposing item difficulty into theory-driven, pre-specified predictors - such as propositional density, syntactic complexity, lexical difficulty, or inferencing demands - thus providing a transparent, replicable rationale for why an item is difficult. By linking each item to its hypothesized sources of difficulty through a Q-matrix, LLTM enables test developers to explicitly align construct definition, item design, and statistical modeling (Kubinger, 2009). This alignment is particularly critical in reading comprehension tests, where difficulty typically stems from a combination of complex, multilayered cognitive operations rather than a single latent trait.

Despite its theoretical appeal, empirical applications of LLTM in language testing - especially in L2 and EFL contexts—remain limited. Most existing studies have focused on native-speaking populations (Brizuela & Montero-Rojas, 2013; Rahman et al., 2022; e.g., Sonnleitner, 2008), with only a few examining LLTM in foreign language proficiency testing. For example, Baghaei and Ravand (2016) is one of the few LLTM studies situated in an EFL context; they found inferencing demands to be the most cognitively challenging component for advanced learners. However, even this work emphasized model fit and did not fully explore LLTM's potential to enhance fairness and interpretability in item development. Given the distinct cognitive and linguistic challenges faced by L2 readers - including reduced lexical access, limited syntactic automatization, and variable background knowledge - there is a pressing need for LLTM applications that explicitly account for construct-relevant features in test design.

The present study seeks to contribute to this underexplored area by applying LLTM to the reading comprehension subtest of the Foreign Language Assessment Program (PELEx, for its acronym in Spanish). Currently, the PELEx reading comprehension subtest uses Rasch modeling to estimate item difficulty (A. Fallas, personal communication, March 14, 2024). In this one-parameter logistic model, all items are assumed to discriminate equally among test-takers, while the difficulty parameter varies (Embretson & Reise, 2000; Rasch, 1960). Rasch modeling is widely used in large-scale foreign language tests, including ACTFL's Reading Computer Adaptive Test (Tschirner et al., 2013), the Hausa CAT (Dunkel, 1999), the Monash/Melbourne French CAT (Burston & Monville-Burston, 1995), the R-CARPE Russian test (Larson, 1999), and SIMTEST for English (Sumbling et al., 2007). Its strengths include comparability, scalability, and adaptability in computer-based testing. However, Rasch's reliance on response data alone limits the extent to which item difficulty can be interpreted by stakeholders - particularly educators, curriculum developers, and learners.

To address this limitation, the present study proposes an LLTM-based reanalysis of the PELEx reading comprehension subtest, using a Q-matrix that encodes a set of theoretically informed predictors such as text length, syntactic complexity, and inference type. Rather than relying solely on statistical fit, this approach enables a more nuanced examination of what makes reading comprehension items more or less difficult - thus offering a more transparent and actionable framework for item calibration.

To guide this investigation, the study addresses the following research question: To what extent can the pre-specified predictors included in the Q-matrix explain item difficulty in the PELEx reading comprehension subtest using the Linear Logistic Test Model (LLTM)?

The study contributes to the growing body of scholarship that advocates for psychometric models grounded not only in statistical rigor but also in cognitive theory and practical relevance. This

contribution is especially timely given the increasing demand for fairness, transparency, and interpretability in digital, large-scale, and high-stakes language testing environments.

Literature Review

Commonly Used Models in Foreign Language Testing

Foreign language testing has historically relied on psychometric models to estimate test-taker ability and item difficulty. Among these, Item Response Theory (IRT) has emerged as the dominant framework, offering advantages over Classical Test Theory (CTT) by providing item-level information and accounting for measurement precision at different ability levels (Embretson & Reise, 2000).

The Rasch model (Rasch, 1960) is widely used in language assessment, providing valuable insights into test fairness, design, and evaluation (Fan & Knoch, 2019). As a one-parameter logistic model (1PL), Rasch assumes that item difficulty is the sole determinant of performance, treating item discrimination as a constant across all items. The Rasch model's key strength is its property of specific objectivity, ensuring that item difficulty and test-taker ability estimates remain independent. This characteristic facilitates test equating and adaptive testing, making Rasch a preferred model in high-stakes language assessments such as TOEFL and PISA (Aryadoust et al., 2021). However, the equal discrimination assumption limits its flexibility, making it less suitable for modeling complex language constructs with diverse item types (Fan & Bond, 2019).

To address Rasch's limitations, the Two-Parameter Logistic (2PL) model introduces item discrimination as an additional parameter, allowing items to contribute differently to ability estimation. This model is particularly useful in reading and listening assessments, where some items inherently differentiate between high- and low-ability test-takers more effectively (Debelak & Strobl, 2019; Embretson & Reise, 2000). However, the 2PL model requires larger sample sizes for stable parameter estimation, making it less practical for small-scale assessments.

The Three-Parameter Logistic (3PL) model (Birnbaum, 1968) further extends the 2PL model by incorporating a guessing parameter, accounting for the probability that test-takers answer correctly by chance. The 3PL model is particularly relevant in multiple-choice assessments, such as standardized English proficiency tests, where low-ability test-takers may guess correctly (Hambleton & Swaminathan, 1985). However, estimating the guessing parameter requires large datasets, and some researchers argue that it can lead to overfitting, reducing interpretability (Embretson & Reise, 2000; Paek et al., 2023).

Although these IRT models provide robust statistical foundations for foreign language testing, they remain difficult to interpret for non-technical stakeholders such as test-takers, teachers, and policymakers. The Rasch model's single difficulty parameter is relatively intuitive, but the additional parameters in 2PL and 3PL introduce complexity that can make item difficulty harder to understand. This lack of transparency is particularly problematic in educational settings, where teachers and students rely on difficulty estimates to guide instruction (Messick, 1995). In foreign language testing contexts specifically, this interpretability challenge is compounded by a mismatch between the logit scales on which Rasch difficulty parameters are reported and the proficiency frameworks, such as the CEFR, that are more familiar to teachers, curriculum designers, and test-takers (McNamara & Ryan, 2011; Pill & McNamara, 2016). Where stakeholders lack the psychometric training to interpret item difficulty in logit units, technically sound Rasch outputs may offer limited practical guidance for instructional or developmental decisions.

The Role of LLTM in Addressing Transparency

The Linear Logistic Test Model (LLTM) (Fischer, 1973) was introduced to enhance the interpretability of item difficulty estimation, addressing a key limitation of other Item Response Theory (IRT) models. Unlike the models presented before, LLTM explicitly decomposes difficulty into measurable cognitive predictors, allowing for a theory-driven analysis of test performance (Baghaei & Kubinger, 2015; Embretson & Daniel, 2008).

This cognitive predictor decomposition is achieved through the use of a Q-matrix, which maps test items to underlying cognitive predictors (Tatsuoka, 1983). Unlike IRT models that estimate item difficulty purely from response data, LLTM links item complexity directly to theoretical cognitive attributes (Jang, 2009). Each row in the Q-matrix represents a test item, while each column corresponds to a specific cognitive or linguistic feature influencing difficulty. The mathematical foundation of LLTM represents item difficulty as the sum of weighted cognitive predictors, ensuring that test difficulty is explicitly linked to theoretically defined cognitive properties (Kubinger, 2009).

By integrating item-design features into the estimation of item difficulty, LLTM provides a more transparent and interpretable framework that facilitates principled test construction and fairness (Baghaei & Hohensinn, 2017; De Boeck & Wilson, 2004; Fischer, 1973; Kubinger, 2009). For example, in language testing, a Q-matrix enables researchers to systematically evaluate how linguistic structures, inferencing demands, and working memory constraints may contribute to item complexity, which in turn can help stakeholders understand with clear justifications what specific elements drive item difficulty on a reading or listening comprehension test (Sawaki, et al., 2009).

In addition to interpretability, LLTM supports fairness in assessment by systematically modeling item difficulty based on explicit cognitive predictors, thereby reducing construct-irrelevant variance (Graßhoff et al., 2010). Traditional psychometric models infer difficulty post hoc, often leading to unexpected bias, whereas LLTM allows for prior calibration, ensuring that difficulty variations are tied to theoretically relevant variables (Baghaei & Hohensinn, 2017). This structured approach enhances test fairness across diverse learner populations by providing evidence-based difficulty estimations rather than relying solely on statistical abstractions.

However, LLTM presents several challenges. Its reliance on the Q-matrix framework introduces potential subjectivity in defining predictors, making misclassification errors a risk (Kubinger, 2009). Additionally, the additivity assumption, which models difficulty as a linear function of cognitive predictors, may fail to capture interaction effects between linguistic features, potentially oversimplifying predictor complexity (Fischer, 1973). Some empirical studies have also noted poorer model fit in certain LLTM applications compared to Rasch-based models, suggesting that parameter constraints may affect predictive validity (Alexandrowicz, 2011).

Despite these limitations, recent research has explored adaptations of LLTM, including Bayesian modeling approaches that adjust for within-test learning effects and dynamic item weighting in adaptive testing environments (Lozano & Revuelta, 2023). These extensions aim to improve parameter estimation efficiency and enhance LLTM's applicability in large-scale assessments. As LLTM continues to evolve, its role in transparent, theory-driven assessment design remains an important area for further exploration.

Applications of LLTM in Reading Comprehension Tests

Empirical research has demonstrated that key linguistic predictors—such as propositional density, syntactic complexity, vocabulary difficulty, and inferencing demands—significantly impact test difficulty (Baghaei & Ravand, 2016; Brizuela & Montero-Rojas, 2013; Sonnleitner, 2008). Among

linguistic predictors, propositional density has emerged as a major determinant of reading comprehension difficulty, referring to the number of distinct semantic units or meaning-bearing propositions within a given text (Brizuela & Montero-Rojas, 2013). A proposition, in this context, represents a fundamental idea or a relationship between concepts, typically comprising a subject and a predicate. Texts with high propositional density encode multiple interrelated ideas, requiring readers to process and integrate various pieces of information simultaneously, thereby increasing cognitive load and comprehension difficulty (Brizuela & Montero-Rojas, 2013; Kintsch, 1998).

Syntactic complexity is another critical factor, as grammatical structures influence the cognitive effort required for sentence parsing. Items featuring passive voice, negations, and multiple embedded clauses impose additional processing demands, particularly for non-native speakers who must engage in both syntactic parsing and lexical access (Baghaei & Ravand, 2016; Just & Carpenter, 1992). Research shows that long-distance dependencies and center-embedded clauses contribute to increased response times and error rates in comprehension assessments (Gibson, 2000; Grodner & Gibson, 2005). The effects are particularly pronounced among L2 learners, who must simultaneously process syntax while accessing lexical meaning, amplifying difficulty (Ellis, 2006). From a test design perspective, controlling unintended syntactic complexity is crucial to ensuring that difficulty reflects reading comprehension ability rather than linguistic barriers.

Beyond syntactic and propositional complexity, lexical difficulty is another predictor of comprehension performance. Lexical complexity—measured by word frequency, word length, and morphological transparency—directly affects word recognition speed, semantic retrieval, and overall comprehension efficiency (Nation, 2001). Lexical factors, such as low-frequency words and specialized vocabulary, contribute to item difficulty in reading comprehension assessments (Brizuela & Montero-Rojas, 2013). Baghaei & Ravand (2016) identified vocabulary knowledge as one of the five cognitive processes involved in reading comprehension assessments. However, their findings indicate that vocabulary was the least challenging process for advanced English as a foreign language (EFL) learners, whereas inference-making posed the greatest difficulty. This suggests that while lexical knowledge plays a role in comprehension, higher-order cognitive processes, particularly inference-making, impose a significantly greater cognitive demand in advanced reading assessments.

Another key determinant of reading comprehension difficulty is inference-making, which imposes additional cognitive strain by requiring test-takers to integrate prior knowledge with textual information—a process central to the Construction-Integration Model (Kintsch, 1998). Unlike direct retrieval tasks, inference-based questions require readers to construct bridging inferences (establishing coherence across sentences) and elaborative inferences (supplementing missing details) (Grabe, 2008). These inferential processes increase working memory load and are particularly challenging for second-language learners, who often face additional constraints such as limited lexical access, weaker syntactic processing, and reduced background knowledge.

Beyond these core predictors, some research has also examined multimodal influences—such as the integration of visual representations or audio support—on reading comprehension difficulty (Krell et al., 2021; Nushi & Jahanbin, 2024), though these factors remain peripheral to the primary linguistic and cognitive demands that are central to LLTM-based frameworks for item difficulty modeling.

Given the complexity of reading comprehension and the wide range of linguistic and cognitive factors affecting item difficulty, LLTM provides a robust framework for systematically modeling these influences. However, while LLTM has been widely used in L1 reading research, its application to L2/EFL contexts remains extremely limited. Among the studies reviewed, Baghaei and Ravand (2016) applied LLTM in an EFL context, investigating how inferencing, main idea identification, syntactic complexity, vocabulary knowledge, and reading for details influence item difficulty in the

Iranian National University English Exam (INUEE). Their findings indicated that inferencing was the strongest predictor of difficulty, while vocabulary was the least challenging, with their LLTM model explaining 56% of the variance ($\Delta^2 = 0.56$) in item difficulty.

By contrast, most LLTM-based studies have focused on L1 reading comprehension. Brizuela and Montero-Rojas (2013) investigated Spanish L1 readers using LLTM and reported that their models explained between 51% and 85% of item difficulty variance ($r^2 = 0.51–0.85$). However, their study did not isolate the exact contribution of propositional density and syntactic complexity within this total variance explained. Similarly, Rahman et al. (2022) examined LLTM in English L1 reading assessments, identifying text genre, inferencing, and task interactions as major contributors, with 38% of variance explained ($\Delta^2 = 0.38$).

Although these studies confirm that LLTM effectively models reading comprehension difficulty, LLTM applications have not systematically explored how predictor influences differ between L1 and L2 readers. Unlike L1 readers, L2/EFL learners must engage in explicit lexical retrieval, syntactic decoding, and additional working memory processing due to lower automaticity in language processing (Grabe, 2008). Consequently, predictor weights in LLTM models may differ significantly between L1 and L2 populations, yet existing research has not systematically explored these differences.

Taken together, the studies reviewed reveal two important and related gaps in the LLTM literature. First, empirical applications of LLTM in L2 and EFL reading contexts remain strikingly limited. With the notable exception of Baghaei and Ravand (2016), existing work has focused predominantly on L1 populations, leaving largely unanswered the question of how predictor weights may differ when readers must simultaneously manage reduced lexical automaticity, constrained syntactic processing, and variable background knowledge (Grabe, 2008). Because these processing constraints are constitutive of L2 reading rather than incidental to it, L1-derived LLTM models cannot be assumed to generalize to foreign language assessment contexts. Second, even the few EFL-based applications have prioritized model fit over the practical implications of LLTM for test development. The potential of cognitive predictor decomposition to make item difficulty more transparent, to support principled item writing, and to provide a basis for fairness reporting to non-technical stakeholders has remained largely unexplored (Baghaei & Hohensinn, 2017; Kubinger, 2009).

The present study addresses both gaps by applying LLTM to the PELEx English proficiency reading comprehension subtest, a large-scale certification exam administered to EFL learners in Costa Rica. Using a theoretically grounded Q-matrix that encodes cognitive and linguistic predictors drawn from the reading comprehension literature, this study aims not only to examine the extent to which those predictors account for item difficulty in this EFL context but also to demonstrate how the LLTM framework can serve as a practical tool for improving the transparency and fairness of item difficulty estimation in foreign language test development.

Methods

This study adhered to a systematic protocol for item difficulty modeling, following the framework outlined by Embretson and Daniel (2008). The process comprised five key steps: developing a conceptual foundation for item complexity, obtaining relevant item response data, scoring item stimulus features to represent cognitive complexity, selecting a statistical method, and interpreting results.

Developing a Conceptual Foundation for Item Complexity

The Q-matrix for this study was adapted from the cognitive model for item complexity in Embretson and Daniel (2008) and developed based on the PELEx test blueprint, which aligns with CEFR guidelines for reading comprehension (Araya Garita et al., 2022). Predictors were selected to reflect item features hypothesized to influence difficulty, such as text length, syntactic complexity, and the presence of inferential reasoning. These predictors were operationalized and categorized into 16 cognitive and linguistic dimensions. The process of finalizing the Q-matrix involved two separate inter-rater calibration exercises conducted by the two researchers, who are TESOL (Teaching English as a Second Language) experts, using 19 random items from the item bank. This iterative process was challenging and required revisiting and refining the definitions of predictors to ensure clarity and alignment with the conceptual framework and item design criteria. Key adjustments included:

- The predictor "Reading strategies" was removed because it overlapped entirely with "Comprehension knowledge," rendering it redundant.
- "Recall language rules" was excluded, as it was deemed to represent overall proficiency rather than a specific item characteristic.
- The "Number of inferences" predictor was reframed as "Presence of inferences needed" to enhance precision.
- Using one word-count processor consistently made a huge improvement. When using ChatGPT, Microsoft Word Online, and Microsoft Word Desktop, all those three produced different word counts, needed for the predictors "Text complexity" and "Context". The researchers opted for Microsoft Word Desktop.

These refinements ensured the Q-matrix captured the essential item features relevant to the study. Table A1 (Supplemental materials) presents the final set of 14 predictors and their operationalizations.

Obtaining Relevant Item Response Data

Data for this study were drawn from the reading comprehension subtest of the PELEx English language certification test administered in 2024. This subtest comprised 50 multiple-choice items graded dichotomously (correct/incorrect) and completed by adult 878 test-takers. The construct measured is understanding of non-technical English related to both regional and global contexts that pertain to the socio-interpersonal, transactional, and academic domains, formally and informally while using as reference the descriptors of CEFR (Araya Garita et al., 2022). The items were randomly sampled from a calibrated item bank, with 10 items targeting each CEFR proficiency level (A1–C1). This de-identified secondary dataset was provided by PELEx; hence, there was no need for an IRB approval for conducting this study.

Scoring Item Stimulus Features to Represent Cognitive Complexity

The finalized Q-matrix was used to systematically score only 50 items in the reading comprehension subtest. This scoring process involved evaluating each item against 14 pre-specified predictors hypothesized to influence item difficulty. Scoring was conducted item-by-item, and the resulting dataset captured the full range of linguistic and cognitive dimensions for each of the 50 items. The Q-matrix predictors were operationalized consistently across items, which provides the definitions and examples for each predictor. This systematic approach to scoring ensured that the item features were objectively quantified and aligned with the theoretical framework, forming the basis for subsequent statistical analysis using the Linear Logistic Test Model (LLTM).

To ensure reliability in the scoring, the two researchers independently coded the items on the 14 Q-matrix predictors and then computed inter-rater agreement using Cohen's kappa. The unweighted kappa was 0.68, 95% CI [0.63, 0.72], indicating substantial agreement (Landis & Koch, 1977). After this initial independent coding, discrepancies were reviewed and resolved to produce a single final Q-matrix, which was then used in all LLTM analyses.

Selecting a Statistical Method

The Rasch model was used to establish baseline item difficulty parameters, which was also the model employed by PELEx to calibrate their items (see Supplemental materials for a discussion on model fit statistics and IRT assumptions for this test). This model estimates the probability of a correct response using the equation:

$$P(X_{is} = 1) = \frac{e^{\theta_s - \beta_i}}{1 + e^{\theta_s - \beta_i}}$$

where $(P(X_{is} = 1))$ represents the probability of a correct response by examinee (s) to item (i), (θ_s) is the examinee's trait level, and (β_i) is the item's difficulty parameter.

The Linear Logistic Test Model (LLTM) was then applied to determine whether the pre-specified predictors in the Q-matrix accounted for item difficulty. The LLTM extends the Rasch model by modeling item difficulty $((\beta_i'))$ as a linear combination of predictors:

$$P(X_{is} = 1) = \frac{e^{\theta_s - \beta_i'}}{1 + e^{\theta_s - \beta_i'}}, \quad \beta_i' = \sum_m \eta_m q_{im} + \eta_0$$

where (q_{im}) indicates the score for variable (m) on item i, (η_m) represents the weight of variable (m) on item difficulty, and (η_0) is the baseline constant. By examining the weights (η_m), predictors, such as the presence of visual aids or the need for paraphrasing, test developers can identify how each layer of the cognitive complexity matrix contributes to the construction of their item difficulties.

The LLTM was implemented using the eRm package in R, and model fit was assessed by comparing it to the Rasch model and a null model.

Interpreting Results

Model fit was assessed by comparing the Linear Logistic Test Model (LLTM) with the baseline Rasch model and a null model. According to Embretson (1997), the null model constrains all item parameters to be equal, providing a baseline for evaluating the explanatory power of more complex models. In contrast, the (LLTM) uses pre-specified predictors to explain item difficulty. Comparisons were made using the following fit statistics (Embretson, 1997):

$$\Delta^2 = \frac{-2 \ln L_{\text{null}} - 2 \ln L_{\text{target}}}{-2 \ln L_{\text{null}} - 2 \ln L_{\text{saturated}}}$$

where:

- Saturated model: Includes unique parameters for each item, representing the best possible model fit, such as in the Rasch model.
- Null model: Assumes all item parameters are equal, serving as the simplest possible model.
- Target model: Represents the LLTM, which incorporates predictors from the Q-matrix.

A significant chi-square test $\left(\chi^2 = -2 \ln L_{\text{null}} - 2 \ln L_{\text{target}}, p < 0.05\right)$ indicates that the LLTM provides a better fit than the null model. This test evaluates whether the inclusion of predictors in the LLTM improves model performance relative to the assumption of uniform item difficulty.

The Δ^2 fit statistic quantifies the improvement in model fit as predictors are added to explain item difficulty. This will indicate what percentage of the total item difficulty in a test is accounted for by the predictors in the Q-matrix.

Results

Model Comparisons

Table 1 presents a comparison of three models using our clean dataset: the null model, the LLTM, and the Rasch model. The null model, which assumes all items have equal difficulty, served as a baseline for evaluating the fit improvement achieved by incorporating predictors in the LLTM. The Rasch model, which estimates item-specific difficulty parameters, was included to establish a benchmark for item calibration.

Table 1. *Model comparisons*

Model	# Par.	$-2\ln L$	AIC	BIC	χ^2	*df*	Model fit Δ^2
Null	1	34962.82	34964.82	34969.6	-	-	-
LLTM	14	25636.02	25664.02	25730.91	9326.8***	13	0.765
Rasch	47	22771.84	22865.84	23090.39	-	-	1

Note. Significant at the *p* < 0.001 alpha level.

The fit of each model was evaluated using $-2\ln L$, with smaller values indicating better fit. The null model had the poorest fit $-2\ln L = 34{,}962.82$, followed by the LLTM ($-2\ln L = 25636.02$) and the Rasch model ($-2\ln L = 22{,}771.84$).

The chi-square test (χ^2) revealed significant differences between the LLTM and the null model ($\chi^2 = 9326.8, df = 13, p < 0.001$). These results indicate that the predictors included in the LLTM significantly improve the explanation of item difficulty compared to simpler models.

For this study, $\Delta^2 = 0.765$, indicating that 77% of the variance in item difficulty is explained by the LLTM predictors.

Weights and source of complexity

The LLTM identified significant contributions from the predictors specified in the Q-matrix. Table 2 summarizes the weight estimates (η_m) for each predictor, along with their standard errors and 95% confidence intervals.

Table 2. *Predictor weights in LLTM model sorted by weight in descending order*

Predictor	Estimate	Std. Error	Lower CI	Upper CI	Cor. CTT
Subtask complexity	0.654	0.048	0.560	0.749	-0.649***
Inferences	0.445	0.061	0.326	0.564	-0.280
Sentence complexity	0.134	0.011	0.113	0.155	-0.178
Paraphrasing	0.125	0.039	0.048	0.203	-0.361*
Relative definition	0.015	0.049	-0.082	0.111	-0.381*
Text complexity	-0.007	0.001	-0.008	-0.006	-0.391*
Context	-0.027	0.002	-0.030	-0.024	-0.587***
Linguistic complexity	-0.198	0.054	-0.303	-0.093	-0.318
Meaning interpretation	-0.411	0.070	-0.547	-0.274	-0.445*

Grammar in context	-0.759	0.047	-0.850	-0.667	-0.552***
Comprehension knowledge	-0.864	0.039	-0.941	-0.788	-0.714***
Translate visuals	-0.972	0.164	-1.294	-0.650	0.420*
Decision	-1.260	0.056	-1.369	-1.151	-0.588***
Encode visuals	-1.702	0.157	-2.009	-1.394	-0.539***

Note. Cor.CTT indicates the correlation of each predictor with item difficulty p-values computed using Classical Test Theory (Table A2 in Supplemental materials). *significant at the 0.05 alpha level, ** significant at the 0.01 alpha level, *** significant at the 0.001 alpha level.

The predictors with the highest weights in item difficulty—those contributing to making items more difficult—were "Subtask complexity", "Inferences", "Sentence Complexity", and "Paraphrasing". It is noteworthy that most of these predictors assess in their own way test takers' capabilities to move beyond the literal words on the page to endorse an item, requiring them to imply or paraphrase concepts, interpret meaning, discern intent, or capture tone. In essence, the presence of implicitness and the need for text rewording appear to be the primary contributors to item difficulty for this population taking a reading proficiency test. Figure 1 best illustrates how these predictors' weights contribute to this test's item difficulty (in green).

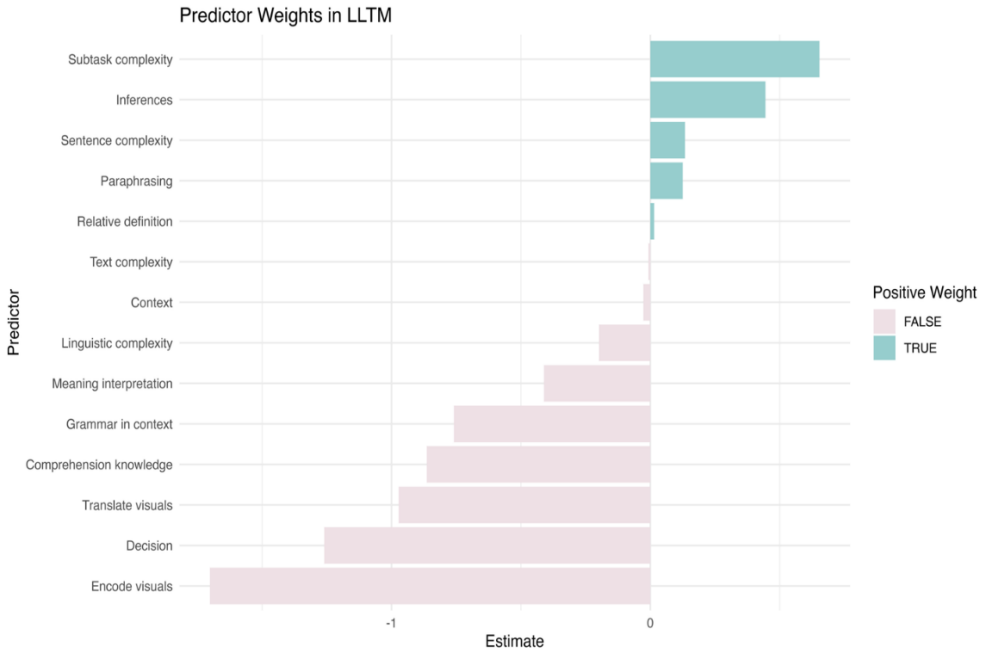


Figure 1. *Horizontal Bar Chart of Predictor Weights in LLTM*

The predictors with the lowest weights—indicating a weaker contribution to item difficulty—were "Encode visuals," "Decision," "Translate visuals," "Comprehension knowledge", and "Grammar in context" (in lavender in Figure 1). For a test not exclusively focused on the A1 CEFR population, it is expected that predictors like "Grammar in context" and "Translate visuals" are associated with lower item difficulty. Items linked to these predictors often do not require extensive grammatical decoding to arrive at the correct answer and are typically accompanied by visual aids. In such cases, students can rely on scanning for simple words or locating relevant information directly in the text or visuals—a skill that aligns with the basic user competencies described for A1 and A2 levels (Council of Europe, 2020).

What is less expected, however, is that "Comprehension knowledge" has a weaker contribution to item difficulty. This predictor is traditionally considered a key driver in item development and categorization, often guiding item writers in aligning items with specific CEFR bands. If anything, "Comprehension knowledge" was anticipated to have a stronger positive weight, given its fundamental role in reading proficiency tests.

While most predictors exhibited statistically significant contributions to item difficulty, a subset demonstrated weights that were statistically close to zero. These predictors were "Relative definition," "Text complexity," and "Context." Most of these predictors measure structural features, such as the number of sentences or words in the text, item stem, or response options. While theoretically relevant, these predictors appear to have a negligible impact on item difficulty within this test population. These weights may occur because these three predictors are highly correlated (Table A3 in Supplemental materials) with other predictors.

To systematically evaluate whether these predictors could meaningfully contribute to the model, multiple post-hoc transformations were applied. First, a composite score averaging "Text Complexity" and "Sentence Complexity" was created to account for their shared variance ($r = 0.81$), resulting in a new composite weight of -0.001. Second, categorical recoding was tested, segmenting "Text Complexity" into discrete bands based on text length (1= < 100 words, 2= > 100 and < 200, and so on), while eliminating also "Sentence complexity," resulting in a weight of -0.060. Regardless of the approach, results remained consistent, with none of these transformations substantially improving the predictive power of these variables in relation to item difficulty.

LLTM item difficulty coefficients

Based on the predicted item difficulty coefficients obtained from the LLTM (Table A2 in Supplemental materials), the items largely aligned with their intended CEFR categorization by PELEx (Figure 2).

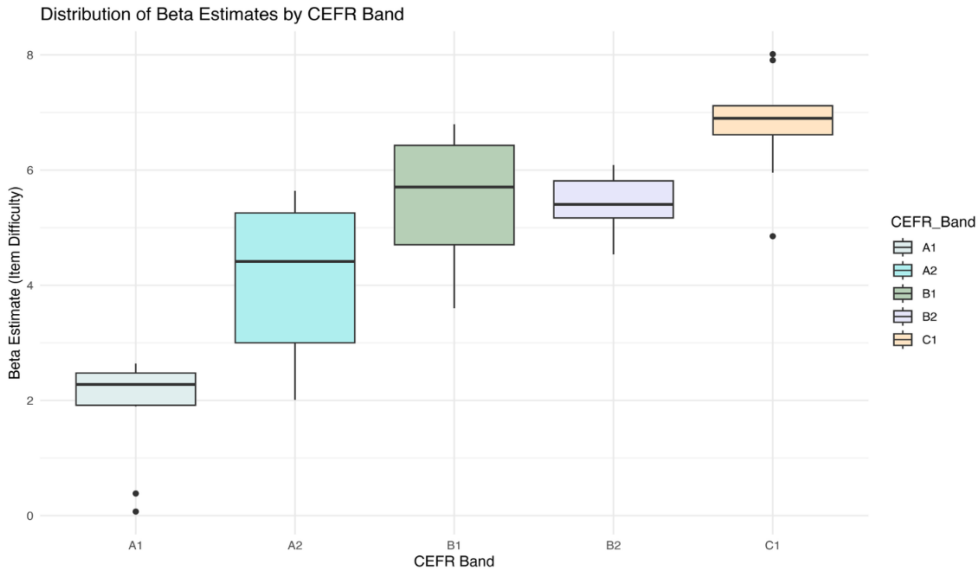


Figure 2. *Box Plots of Item Clusters Based on Difficulty Estimates by CEFR Band*

A1 items exhibited the lowest difficulty values, indicating their accessibility to lower-ability test-takers. Within this band, several items demonstrated near-identical difficulty estimates, suggesting consistency in their cognitive demands.

A2 items followed the A1 cluster but exhibited greater variability in their difficulty estimates. Notably, some A2 items approached the difficulty range of B1 items, which may reflect transitional skills between basic and independent user levels.

B1 and B2 items formed a collective, uniform cluster, with overlapping difficulty estimates, consistent with the CEFR classification of these levels as the two components of the independent user category (Council of Europe, 2020). This pattern indicates a smooth progression in item difficulty across these levels, though slight variability suggests room for refinement in distinguishing between B1 and B2 boundaries.
Finally, C1 items predominantly clustered at the higher end of the difficulty spectrum, confirming their role as the most challenging items in the test. The pronounced gap between C1 items and the preceding B2 cluster highlights the significant cognitive leap required to achieve mastery at the proficient user level. This same item difficulty distribution was observed when estimating item difficulty using Classical Test Theory (Table A2).

Discussion

This study sought to enhance the transparency and fairness of the PELEx English proficiency reading comprehension subtest by applying the Linear Logistic Test Model (LLTM). By aligning item difficulty estimates with pre-specified cognitive and linguistic predictors, the findings provide actionable insights for test developers seeking to refine item design and better communicate the rationale behind item difficulty to stakeholders.

The interpretations that follow should remain provisional. Because the analysis was conducted on a reading form assembled from randomly selected banked items, and because the psychometric evidence for that form was mixed (see Supplemental materials), the LLTM results are better viewed as exploratory evidence for this specific administration than as a definitive explanatory model of item difficulty in the reading subtest as a whole.

Model fit

The LLTM model explained a substantial portion of the variance in item difficulty, as reflected in the Δ^2 value of 0.765. This indicates that 77% of the variance in item difficulty was accounted for by the cognitive and linguistic predictors specified in the Q-matrix. This result is particularly significant for test developers, as it demonstrates that the chosen predictors effectively capture the key dimensions of item complexity, aligning with the findings in LLTM literature (Baghaei & Ravand, 2016; Brizuela & Montero-Rojas, 2013). The high Δ^2 suggests that the Q-matrix captured a substantial share of the item difficulty observed in this test form, offering a practical framework for understanding and communicating item difficulty beyond traditional Rasch-based approaches.

Predictor Weights and Item Complexity

The results suggest that implicitness was an important source of item difficulty in this form. Predictors such as "Paraphrasing," "Subtask complexity," and "Inferences" were among the most significant contributors, indicating that items requiring test-takers to infer, reword, or interpret implicit meaning demand greater cognitive effort. This aligns with the principles of reading proficiency, where higher-order comprehension skills differentiate more advanced users from basic users (Baghaei & Ravand, 2016; Brizuela & Montero-Rojas, 2013; Sonnleitner, 2008).

Conversely, predictors like "Grammar in context" and "Translate visuals" contributed to lower item difficulty, as expected for lower proficiency levels (A1 and A2). These items allow test-takers to rely on simpler decoding strategies, such as locating key words or interpreting visuals without fully engaging with the text. However, the unexpected role of "Comprehension knowledge" in lower

item difficulty challenges traditional assumptions about its function in distinguishing advanced proficiency levels. This finding may stem from the broad nature of "Comprehension knowledge," which encapsulates the general proficiency required to endorse an item. Its strong correlations with other predictors (Table A3) suggest that it may overlap with their operationalization. Since this Q-matrix was applied to a subset of the PELEx item bank, it remains unclear whether this is a sample-specific effect or a broader pattern. Future research should examine whether this predictor exhibits similar behavior with a different combination of items and another sample. If similar patterns emerge, test developers may consider the exclusion of this predictor in future models.

As for the predictors with negligible impact on item difficulty ("Relative definition," "Text complexity," and "Context"), different transformations were tested to assess their potential contribution to the model. These included recoding "Text Complexity" into categorical bands and creating a composite score by combining "Text Complexity" and "Sentence Complexity" due to their strong correlation. Regardless of the approach, results remained consistent, showing that these predictors did not substantially influence item difficulty in this sample. Given their limited explanatory power, future research should examine whether this pattern persists across different samples before considering any modifications to the model.

Alignment with CEFR Bands

The alignment of predicted item difficulty coefficients with CEFR bands provides evidence toward the validity of the LLTM framework. Items designed for A1 levels consistently exhibited the lowest difficulty values, while C1 items were the most challenging. Intermediate bands (B1 and B2) displayed a generally logical progression in difficulty, though some overlap in difficulty estimates is evident. However, such overlap is to be expected, as CEFR levels are inherently nested, and language proficiency is a skill built progressively from simple to more advanced features. The overlap shown in Figure 2 is consistent with the nested nature of CEFR levels and suggests that the LLTM was able to recover a generally plausible progression of difficulty across proficiency bands in this form.

Moreover, these LLTM-based difficulty estimates may still be useful for reviewing how items are categorized within the item bank. By identifying overlaps or inconsistencies in item classification, test developers can improve the alignment of items with their intended CEFR levels, enhancing consistency and transparency in the test design process. In that sense, the approach may offer a useful starting point for a more systematic review of item categorization and item-feature specifications.

Implications for Test Development

For test developers, the main contribution of this study is that it offers a more explicit basis for examining what appears to make reading items difficult in the PELEx item bank. In the present form, the strongest weights were associated with *Subtask complexity*, *Inferences*, and *Paraphrasing*, suggesting that item difficulty was tied less to surface features alone and more to demands requiring readers to go beyond literal retrieval and reconstruct meaning. This is useful for item development because it identifies the kinds of features that may warrant closer attention during item writing, review, and calibration. At the same time, several predictors showed weak, negligible, or unexpected contributions, indicating that parts of the current Q-matrix still require refinement. In that sense, the value of the study is not that it establishes a final explanatory model, but that it provides a structured starting point for revising predictor definitions, improving item specifications, and strengthening the interpretability of item difficulty estimates within the PELEx item bank. This also has implications for fairness. By expressing item difficulty in terms of explicit cognitive and linguistic features, the LLTM offers a more transparent rationale for why some items

may be more demanding than others, which may help reduce reliance on psychometric estimates alone when those results need to be communicated to non-technical stakeholders.

Limitations and Future Directions

Several limitations should be considered. Most importantly, the present study was conducted on a reading test composed of randomly selected items from a larger calibrated item bank rather than on a complete fixed form designed and reviewed as a single measurement instrument. This feature of the test form is central to the interpretation of the findings, because the observed psychometric behavior reflects not only the properties of the individual items, but also the particular configuration in which those items were administered together. As documented in the Supplemental materials, the resulting evidence was mixed: the Rasch model showed near-threshold fit on several global indices and was favored by BIC, which supports its use as a parsimonious framework for the present purposes. Nevertheless, the 2PL provided better overall fit; evidence for unidimensionality was only partial; items were flagged for misfit, and a small number of item pairs showed local dependence. Under these conditions, the LLTM results are best interpreted as exploratory rather than strictly explanatory. More specifically, the predictor weights reported here should be understood as an initial indication of how the proposed Q-matrix may account for item difficulty in this assembled form, not as definitive estimates of the structure of reading comprehension difficulty in the PELEx reading subtest more broadly. Future research should therefore evaluate the same Q-matrix on a complete fixed form and examine whether similar predictor weights, explained variance, and item-feature relations are recovered under more stable measurement conditions.

Under the present conditions, more flexible approaches may be informative as a complement to the LLTM. For example, machine learning models, such as random forests regressions (Breiman, 2001), can accommodate nonlinear relations and interactions among item features without requiring the same restrictive measurement assumptions for Rasch modeling, while still using a Q-matrix. Comparing LLTM results with those from more flexible methods would help clarify whether weak or unstable predictor effects reflect limitations in the current operationalization of the Q-matrix or the constraints imposed by a Rasch-based framework.

Finally, future research should prioritize refinement of the Q-matrix itself. In the present study, some predictors showed limited explanatory value, whereas others appeared conceptually or empirically proximate, suggesting that the current specification may not yet be optimally parsimonious. A useful next step would be to evaluate whether closely related predictors, such as *Text complexity* and *Sentence complexity*, are better represented as a single composite indicator, and whether predictors with negligible weights should be reoperationalized or excluded. Replication is also needed under stronger design conditions.

References

Alexandrowicz, R. W. (2011). Statistical and practical significance of the likelihood ratio test of the linear logistic test model versus the Rasch model. *Educational Research and Evaluation, 17*(5), 335–350. https://doi.org/10.1080/13803611.2011.630522

Araya Garita, W., Elizondo González, J., & González Ramírez, A. (2022). Getting stakeholders acquainted with the rationale behind the construct of the English language proficiency test of the University of Costa Rica for the Ministry of Education of Costa Rica. *Estudios de Lingüística Aplicada, 40*(75), 119–143. https://doi.org/10.22201/enallt.01852647p.2022.75.1013

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Baghaei, P., & Hohensinn, C. (2017). A method of Q-matrix validation for the linear logistic test model. *Frontiers in Psychology, 8*, Article 897. https://doi.org/10.3389/fpsyg.2017.00897

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research, and Evaluation, 20*(1), Article 1. https://doi.org/10.7275/8f33-hz58

Baghaei, P., & Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicológica, 37*(1), 85–104.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brizuela, A., & Montero-Rojas, E. (2013). Prediction of the difficulty level in a standardized reading comprehension test: Contributions from cognitive psychology and psychometrics. *RELIEVE – Revista Electrónica de Investigación y Evaluación Educativa, 19*(2), 1–21. https://doi.org/10.7203/relieve.19.1.3149

Burston, J., & Monville-Burston, M. (1995). Practical design and implementation considerations of a computer adaptive foreign language test: The Monash/Melbourne French CAT. *CALICO Journal, 12*(2), 26–46.

Council of Europe – Common European Framework of Reference for Languages: Companion volume

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment – Companion volume*. Council of Europe.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

Debelak, R., & Strobl, C. (2019). Investigating measurement invariance by means of parameter instability tests for 2PL and 3PL models. *Educational and Psychological Measurement, 79*(2), 385–398. https://doi.org/10.1177/0013164418777784

Dunkel, P. (1999). Research and development of a computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 71–90). Cambridge University Press.

Ellis, R. (2006). *The study of second language acquisition* (2nd ed.). Oxford University Press.

Embretson, S. E. (1997). Multicomponent latent trait models. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 305–322). Springer.

Embretson, S. E., & Daniel, R. C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science, 50*(3), 328–344.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Volume I – Fundamental techniques* (1st ed., pp. 81–100). Routledge. https://doi.org/10.4324/9781315187815-5

Fan, J., & Knoch, U. (2019). Fairness in language assessment: What can the Rasch model offer? *Papers in Language Testing and Assessment, 8*(2), 1–27. https://doi.org/10.58379/jrwg5233

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first Mind Articulation Project Symposium* (pp. 95–126). MIT Press.

Grabe, W. (2008). *Reading in a second language: Moving from theory to practice*. Cambridge University Press.

Graßhoff, U., Holling, H., & Schwabe, R. (2010). Optimal designs for linear logistic test models. In A. Giovagnoli, A. Atkinson, B. Torsney, & C. May (Eds.), *mODa 9 – Advances in model-oriented design and analysis* (pp. 241–250). Physica-Verlag. https://doi.org/10.1007/978-3-7908-2410-0_13

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science, 29*(2), 261–290. https://doi.org/10.1207/s15516709cog0000_7

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer-Nijhoff.

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing, 26*(1), 031-73. https://doi.org/10.1177/0265532208097336

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122–149. https://doi.org/10.1037/0033-295X.99.1.122

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.

Krell, M., Khan, S., & van Driel, J. (2021). Analyzing cognitive demands of a scientific reasoning test using the linear logistic test model (LLTM). *Education Sciences, 11*(9), Article 472. https://doi.org/10.3390/educsci11090472

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*(2), 232–244. https://doi.org/10.1177/0013164408322021

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Larson, J. (1999). Considerations for testing reading proficiency via computer-adaptive testing. In M. Chalhoub-Deville (Ed.), *Issues in computer-adaptive testing of reading proficiency* (pp. 91–106). Cambridge University Press.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Lozano, J. H., & Revuelta, J. (2023). A Bayesian random weights linear logistic test model for within-test practice effects. *Applied Psychological Measurement, 47*(7–8), 443–459. https://doi.org/10.1177/01466216231209752

McNamara, T., & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8*(2), 161–178. https://doi.org/10.1080/15434303.2011.565438

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research, 49*(4), 305–328. https://doi.org/10.1080/00273171.2014.931798

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nushi, M., & Jahanbin, P. (2024). The effect of audio-assisted reading on incidental learning of present perfect by EFL learners. *Open Education Studies, 6*(1), Article 20240043. https://doi.org/10.1515/edu-2024-0043

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*(1), 50–64. https://doi.org/10.1177/01466210022031664

Paek, I., Lin, Z., & Chalmers, R. P. (2023). Investigating confidence intervals of item parameters when some item parameters take priors in the 2PL and 3PL models. *Educational and Psychological Measurement, 83*(2), 375–400. https://doi.org/10.1177/00131644221096431

Pill, J., & McNamara, T. (2016). How much is enough? Involving occupational experts in setting standards on a specific-purpose language test for health professionals. *Language Testing, 33*(2), 217–234. https://doi.org/10.1177/0265532215607402

Rahman, T., Alexander, P. A., & Chae, S. E. (2022). Reader attributes, task attributes, and reading comprehension proficiency: The relation revealed by two analytic approaches. *Reading Psychology, 43*(7), 495–522. https://doi.org/10.1080/02702711.2022.2126044

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207–230. https://doi.org/10.2307/1164676

Sawaki, Y., Kim, H. J., & Gentile, C. (2009). Q-matrix construction: Defining the link between constructs and test items in large-scale reading and listening comprehension assessments. *Language Assessment Quarterly, 6*(3), 190–209. https://doi.org/10.1080/15434300902801917

Sonnleitner, P. (2008). Using the LLTM to evaluate an item-generating system for reading comprehension. *Psychology Science Quarterly, 50*(3), 380–398.

Sumbling, M., Sanz, P., Viladrich, M. C., Doval, E., & Riera, L. (2007). Development of a multiple-component CAT for measuring foreign language proficiency (SIMTEST). In *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing*. GMAC Conference on Computerized Adaptive Testing Proceedings

Tatsuoka, K. K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*(4), 345–354. https://doi.org/10.1111/j.1745-3984.1983.tb00212.x

Tschirner, E., Bärenfänger, O., Kutschera, S., & Möhring, J. (2013). *Validating the ACTFL Listening and Reading Proficiency Computer Adaptive Test (ACTFL L&R CAT): Technical report 2013-US-PUB-4*. ACTFL.

ACTFL Technical Report 2013-US-PUB-4

Supplemental materials

Table A1. *Reading comprehension Q-matrix*

Predictor	Definition
1. Translation	
1.1 Linguistic complexity	Text CEFR level based on (https://www.english.com/gse/teacher-toolkit/user/textanaolyzer/requestdetails) A1= 1, A2= 2, B1= 3, B2=4, C1= 5
1.2 Text complexity	Total number of words in text
1.3 Sentence complexity	Total number of sentences in text (complete sentences)
1.4 Context	Total number of words in the stem and answer options.
1.5 Encode Visuals	Indicates the presence of visual elements such as images, charts, or diagrams in the reading material that aid comprehension, excluding pure text. 0= Yes, 1= No
2. Integration	
2.1 Interpretation of Meaning	Indicates if the examinee must interpret implied meaning, such as inference of tone, intent, or understanding of figurative language (e.g., metaphors, irony). 1= Yes, 0= No
2.2 Grammar in Context	Indicates if the examinee must apply grammatical rules in the context of the reading passage to understand meaning (e.g., tenses, conditionals). 1= Yes, 0= No
2.3 Generate Paraphrase or Synonyms	Indicates if the examinee must produce paraphrased versions or select synonyms to answer the item or comprehend the text. 1= Yes, 0= No

2.4 Translate Visual Information	Indicates if the examinee must interpret visual cues (e.g., a chart or map) to comprehend the associated text. 1= Yes, 0= No
3. Solution Planning	
3.1 Subtask Complexity	Subtask complexity present in question. 1= Identify a specific detail (e.g., name, date, place). 2= Identify a detail and recognize basic vocabulary in context. 3= Identify the main idea, organize details, and understand a sequence of events. 4= Identify relationships (e.g., cause-effect), infer meaning, and connect ideas. 5= Synthesize across paragraphs, interpret cultural references, and evaluate arguments. A1= 1, A2= 2, B1= 3, B2=4, C1= 5
3.2 Relative Definition	Indicates if a word or phrase (needed to endorse the item) is defined relative to another concept or context-specific term within the text. 1= Yes, 0= No
4. Solution Execution	
4.1 Comprehension Knowledge	Maximum level of comprehension knowledge required, including understanding of intermediate or advanced structures (A1–C2 according to CEFR). Includes the use of reading strategies: 5= C1: Using cues to infer attitude, mood, or intentions 4= B2: Identifying complex main points 3= B1: Following a sequence of events or identifying simple main points 2= A2: Identifying specific details 1= A1: Identify simple words with the help of familiar and simple subjects
4.3 Presence of Inferences	Indicates if inferences are needed to understand implied ideas, context shifts, or cultural nuances. 1= Yes, 0= No
5. Decision	
5.1 Decision Processing	Indicates if information found in distractors or alternative text parts is necessary to eliminate options or answer the item.1= Yes, 0= No

Table A2. *Predicted item difficulty using LLTM sorted by difficulty estimate in ascending order*

Beta	CEFR band	Estimate	Std. Error	Lower CI	Upper CI	CTT difficulty p-values
beta 3	A1	2.276	0.199	1.885	2.666	0.975
beta 4	A1	2.276	0.199	1.885	2.666	0.986
beta 5	A1	1.896	0.175	1.553	2.238	0.986
beta 6	A1	1.978	0.175	1.636	2.320	0.995
beta 1	A1	2.561	0.236	2.097	3.024	0.918
beta 7	A1	2.414	0.234	1.954	2.873	0.986
beta 2	A1	2.642	0.236	2.180	3.105	0.990
beta 8	A1	2.496	0.234	2.037	2.954	0.997
beta 16	A2	2.011	0.197	1.626	2.397	0.981
beta 10	A1	0.071	0.192	-0.305	0.447	0.992
beta 9	A1	0.385	0.240	-0.085	0.856	0.997
beta 20	A2	2.394	0.166	2.070	2.719	0.984
beta 15	A2	3.162	0.179	2.811	3.514	0.990
beta 14	A2	2.947	0.208	2.540	3.355	0.994
beta 19	A2	4.661	0.254	4.163	5.160	0.851
beta 33	B2	4.804	0.223	4.368	5.240	0.954
beta 35	B2	4.537	0.244	4.059	5.015	0.803
beta 32	B2	6.086	0.241	5.615	6.558	0.759
beta 12	A2	5.400	0.221	4.966	5.834	0.869
beta 30	B1	3.601	0.247	3.117	4.085	0.890

beta 31	B2	6.003	0.234	5.544	6.461	0.876
beta 37	B2	5.168	0.239	4.699	5.637	0.836
beta 17	A2	5.640	0.230	5.189	6.091	0.605
beta 11	A2	4.165	0.210	3.753	4.577	0.953
beta 18	A2	5.342	0.249	4.854	5.830	0.912
beta 24	B1	4.579	0.229	4.130	5.028	0.942
beta 40	B2	5.404	0.242	4.930	5.879	0.767
beta 26	B1	6.226	0.230	5.776	6.676	0.844
beta 25	B1	4.517	0.244	4.040	4.995	0.983
beta 27	B1	5.075	0.240	4.604	5.545	0.880
beta 28	B1	6.495	0.235	6.034	6.956	0.744
beta 42	C1	4.851	0.235	4.389	5.312	0.876
beta 29	B1	6.795	0.234	6.336	7.253	0.510
beta 13	A2	4.991	0.232	4.536	5.446	0.987
beta 49	C1	6.612	0.237	6.148	7.077	0.616
beta 34	B2	5.812	0.253	5.315	6.308	0.893
beta 36	B2	5.716	0.237	5.252	6.180	0.576
beta 45	C1	5.955	0.225	5.514	6.397	0.659
beta 23	B1	5.597	0.229	5.148	6.047	0.508
beta 39	B2	5.299	0.236	4.837	5.760	0.905
beta 21	B1	5.813	0.250	5.324	6.303	0.551
beta 22	B1	6.603	0.246	6.120	7.086	0.697
beta 43	C1	7.116	0.237	6.651	7.582	0.499
beta 48	C1	7.095	0.230	6.644	7.546	0.568
beta 44	C1	6.897	0.231	6.444	7.351	0.383
beta 50	C1	7.906	0.248	7.421	8.391	0.351
beta 47	C1	6.641	0.239	6.174	7.109	0.336
beta 46	C1	8.011	0.220	7.580	8.442	0.190

Note. "Estimate" in this table refers to item difficulty, calculated as the negative of item easiness. In eRm's LLTM, item parameters are derived from a linear combination of predictors rather than being freely estimated or centered around zero. Coefficients for items 38 and 41 are not reported, as no student endorsed those items. Estimates reflected similar patterns in item difficulty using Classical Test Theory, as seen in the last column, where higher values correspond to lower item difficulty.

Table A3. *Correlation matrix among predictors in Q-matrix*

	Lin_Comp	Text_Comp	Sent_Comp	Cont.	Enc_V	M_Inter	Gram.	Paraphr.	Trans_Vis	Sub_comp	Rel_def	Comp_know	Infer.	Decs.
Lin_Comp	1													
Text_Comp	-0.159	1												
Sent_Comp	-0.386	0.811**	1											
Cont.	0.384	0.382	0.137	1										
Enc_V	0.188	0.498*	0.412	0.394	1									
M_Inter	0.382	0.144	-0.010	0.237	0.346	1								
Gram.	0.079	0.493*	0.464	-0.312	0.387	0.191	1							
Paraphr.	0.314	0.292	0.206	0.467	0.342	0.330	0.200	1						
Trans_Vis	-0.092	-0.493*	-0.500*	-0.355	-0.809***	-0.280	-0.461	-0.344	1					
Sub_comp	0.579**	0.277	0.074	0.558**	0.612***	0.675***	0.407	0.546**	-0.537**	1				
Rel_def	0.043	0.151	0.239	0.293	0.334	0.461	0.480*	0.263	-0.241	0.386	1			
Comp_know	0.480*	0.411	0.198	0.496*	0.751***	0.650***	0.468	0.491*	-0.655***	0.903***	0.421	1		
Infer.	0.140	0.044	-0.044	0.144	0.203	0.588***	0.204	0.255	-0.164	0.456	0.304	0.433	1	

Decs.	0.456	0.169	0.106	0.475*	0.557**	0.205	0.338	-0.526**	-0.471*	-0.614***	0.238	-0.646**	0.244	1

Note. Correlations represent Pearson correlation coefficients. Significance levels reflect p-values adjusted using the Holm method for multiple comparisons: ***$p < .001$, **$p < .01$, $p < .05$.

Model fit and IRT assumptions

The reading test analyzed in this study was not a fixed linear form designed and reviewed as one complete instrument. Instead, it consisted of items randomly selected from a calibrated item bank for this administration. This distinction is important for interpreting the psychometric results reported below, because indices such as global fit, item fit, dimensionality, and local dependence reflect not only the behavior of individual items, but also the particular combination of items that appeared together in this form. Accordingly, the results in this section should be interpreted as evidence about the functioning of this assembled form rather than as definitive evidence about the reading subtest in all administrations.

This study used the Rasch model as the primary framework for analysis, consistent with the operational approach adopted by the test developer, PELEx. This alignment supports comparability with their calibration procedures. At the same time, model fit comparisons indicated that the data were not fully captured by a strict Rasch specification. As shown in Tables A4 and A5, the 2PL model demonstrated stronger global fit across all indices and a statistically significant improvement over the Rasch model. Even so, the Rasch model remained defensible for the present study because several fit values were close to commonly used thresholds and the BIC favored the more parsimonious model. For that reason, the Rasch framework was retained for this exploratory LLTM application, while the 2PL results are reported here to document the extent to which a more flexible model improved fit.

Table A4. *Global fit indices for Rasch and 2PL models*

Model	CFI	TLI	RMSEA	SRMSR
Rasch	0.879	0.879	0.038	0.074
2PL	0.996	0.996	0.006	0.046

Note. Adequate fit values based on Maydeu-Olivares & Joe (2014): CFI and TL1 > 0.90; RMSEA < 0.08; SRMSR < 0.05.

Table A5. *Likelihood ratio test results for Rasch and 2PL models*

Model	AIC	SABIC	BIC	logLink	χ^2	df	p
Rasch	28155.80	28234.29	28389.90	-14028.90	-	-	-
2PL	27605.06	27912.62	28522.37	-13610.53	836.737	143	0

The assumptions most relevant to the Rasch model - unidimensionality, item fit, and local independence - were also examined. Unidimensionality of the reading test, intended to assess overall reading comprehension proficiency, was evaluated using Principal Component Analysis (PCA) and IRT Exploratory Factor Analysis (EFA). The first eigenvalue explained 12.05% of the variance, below the 20% threshold suggested by Reckase (1979). In addition, the ratio of the first to the second eigenvalue was 2.35, below the 3:1 criterion recommended by Lord (1980). Complementary IRT EFA results using the 2PL model, which allows distinct discrimination parameters for factor analysis (De Ayala, 2009), showed that the first factor explained 32.3% of the variance. However, when the one-factor and two-factor solutions were compared using information criteria (Table A6), the one-factor model showed lower BIC and SABIC values, favoring the more parsimonious solution despite the better fit of the two-factor model. These results suggest that a dominant general reading factor was present, but the evidence for strict unidimensionality was limited.

Table A6. *Model Information Criteria for Unidimensional and Multidimensional 2PL models*

Model	AIC	SABIC	BIC	logLink	χ^2	df	p
1-Factor	27503.39	27657.17	27962.04	-13655.70	-	-	-
2-Factor	27483.04	27712.11	28166.24	-13598.52	114.352	47	0

Item fit was examined using the S-X² statistic (Orlando & Thissen, 2000). Of the final 48 items analyzed, 24 (50%) were flagged for misfit. This level of misfit should not be interpreted automatically as evidence of poor item quality. Because the form was assembled from randomly selected items in a larger bank rather than from a fixed test reviewed as a single unit, some misfit may reflect the combination of items included in this administration. In other words, some items may function differently depending on the other items with which they appear, especially when multiple passages, CEFR levels, or subskills are sampled together in a single form.

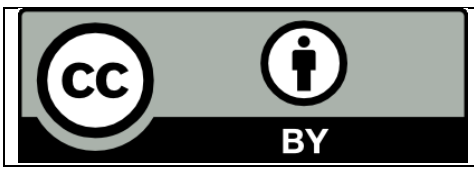
Finally, to evaluate item local independence, residual correlations were examined from the residual matrix produced by the (M_2) statistic, where residuals (≥ 0.20) would be flagged as indicating dependence. 15 pairs out of 1128 (less than 1%) were identified as locally dependent. This percentage of observed dependence may stem from several factors. First, content overlap between items could introduce dependencies, particularly if multiple items assess highly similar subskills or CEFR bands. Second, dependencies may also reflect the way the form was assembled, particularly when items drawn from the same passage or closely related content appear together in a given administration, as it was in case for items 7 and 8 or items 11 and 12.

Overall, these results indicate that the Rasch model provided a workable but not unambiguous framework for this assembled reading form. The evidence was mixed: some indices supported the use of a parsimonious Rasch-based approach, whereas others pointed to limitations in fit, dimensionality, and independence. For that reason, the psychometric evidence reported here is best understood as contextual information for interpreting the LLTM results in this administration, rather than as a definitive statement about the measurement properties of the reading subtest as a whole.

Biographical notes:

Jose Fabián Elizondo González is an instructor and researcher at Universidad de Costa Rica. He recently graduated with his Ph.D. in Educational Psychology at the University of Kansas, USA.

Peyman Jahanbin is a Ph.D. student in Curriculum and Instruction, specializing in TESOL, at the University of Kansas, where he also serves as a Graduate Teaching Assistant in the Department.



Submitted/Received	Accepted
30 January 2026	6 May 2026