Language Testing in Asia

RESEARCH Open Access

Validation of an elicited Imitation test as a measure of Korean language proficiency

Hojung Kim[1*], Changkyung Song[1], Jiyoung Kim[1], Hyeyun Jeong[1] and Jisoo Park[1]

*Correspondence:
renata88@snu.ac.kr

[1] Department of Korean Language Education, Korean Language Education Research Institute, Seoul National University, 1 Gwanak-Ro, Gwanak-Gu, Seoul 08826, Korea

Abstract

This study presents a modified version of the Korean Elicited Imitation (EI) test, designed to resemble natural spoken language, and validates its reliability as a measure of proficiency. The study assesses the correlation between average test scores and Test of Proficiency in Korean (TOPIK) levels, examining score distributions among beginner, intermediate, and advanced learner groups. Using item response theory (IRT), the study explores the influence of four key facets—learners, items, raters, and constructs—on performance evaluation. An explanatory item response modeling (EIRM) analysis identified linguistic factors impacting the EI test's performance. Notably, the study uncovered a robust positive correlation between EI test scores and TOPIK levels. Significant score disparities were observed between beginner and intermediate, as well as beginner and advanced, learner cohorts. An IRT-based exploration of each facet revealed that item difficulty was comparatively lower in contrast to learners' commendable performance, and raters exhibited a high degree of scoring consistency. The EIRM analysis underscores the significance of variables such as the number of syllables, vocabulary score, and content word density in influencing test performance.

Keywords: Elicited imitation, Proficiency, Assessment, Validation, Item response theory

Introduction

Language proficiency is a crucial aspect in second language (L2) acquisition studies as it reflects and engages with the target language. Researchers have extensively explored linguistic attributes and methodological nuances in specific learner groups, considering variables such as learners' first language (L1), linguistic community, age, and language proficiency. Among these, language proficiency stands out as a critical factor, representing learners' capacity to comprehend and engage with the target language. This proficiency parameter serves as a cornerstone for evaluating learners' developmental stages and diagnosing their strengths and areas for improvement (Harsch, 2014, p. 154).

Harnessing the potency of the proficiency variable, which profoundly influences the mastery of language reception and production, requires a valid and dependable means of measuring learners' L2 competence. In the practical context, the efficiency of assessments is paramount, necessitating the choice of assessment methods that seamlessly complement the evaluative process (Han, 2014, pp. 52–53). Consequently, there is a pressing need to conceive and validate assessment tools that are capable

of ensuring evaluation efficiency while providing insights into learners' language competence.

Elicited imitation (EI) tests have gained widespread recognition as a remarkably practical and valid tool for assessing language competence. Their efficacy extends across various dimensions, including item development, assessment implementation, and scoring reliability (Bley-Vroman & Chaudron, 1994; Jessop et al., 2007; West, 2012). In the execution of EI tests, learners are challenged to swiftly comprehend and decode presented sentences, subsequently reconstructing them by employing their own grammatical framework (Wu & Ortega, 2013, p. 683).

A noteworthy advantage of the utilization of EI tests is their ability to offer comprehensive insights into learners' L2 oral proficiency. This is achieved through the assessment of syntactic comprehension and morphological mastery, as reflected in the accuracy of the reproduced sentences (Davis & Norris, 2021; Gaillard & Tremblay, 2016; Vinther, 2002). This distinctive attribute underscores the potential of EI tests to serve as a reliable measure of learners' overall language proficiency in the oral domain.

However, there is a dearth of research on EI tests of Korean as a second or foreign language. The current body of literature is primarily represented by the contributions of Kim et al. (2016) and Isbell and Son (2022). Notably, within the South Korean context, the discussions and investigations surrounding these tests have predominantly centered on their application as assessment tools for children facing language-related challenges (Lee, 2005; Heo & Lee, 2010; Oh & Yim, 2013).

Despite the potential of EI tests, the scope of the related research within the Korean language education and assessment arena remains relatively limited. To address this gap, the present study undertakes a comprehensive validation of a Korean EI test we developed. This process considers the distinctive linguistic attributes of the Korean language and addresses authenticity concerns, drawing inspiration from the EI tests designed by Ortega et al. (2002) and Kim et al. (2016).

The primary focus of our investigation was multi-faceted. Firstly, we explored the correlation between the scores obtained from the EI test we developed and the proficiency levels assessed by the Test of Proficiency in Korean (TOPIK). Additionally, we investigated the variations in EI test score trends across beginner, intermediate, and advanced learner groups. Subsequently, employing the item response theory (IRT), we gauged the difficulty levels of individual items alongside learners' abilities, evaluated the impact of raters, and examined the use patterns of scales and constructs. This analysis enabled us to identify the presence of construct-irrelevant factors.

Furthermore, we also conducted explanatory item response modeling (EIRM) to explore the linguistic factors that influence test performance and item difficulty, encompassing syntactic and lexical aspects that go beyond the number of syllables. A comprehensive exploration such as this is essential for validating the efficacy of EI tests as tools for determining language proficiency.

In light of these objectives, we established the following research questions:

1. How does the developed EI test, employed as a proficiency assessment tool, correlate with TOPIK scores? What trends become evident in EI test scores across distinct TOPIK proficiency levels?

2. What discernible patterns emerge within each facet (learner, item, rater, and construct) of the EI test when subjected to analysis using the IRT?
3. Which linguistic factors exert influence on the performance of the EI test?

Literature review

The utility of EI tests as a measure of proficiency

Since the 1970s, EI tests, which involve the repetition of a heard sentence, have been a staple in the L2 research (Vinther, 2002; Yan et al., 2016). The ease of creating and administering EI tests has rendered them a valuable asset in L2 classrooms and research settings. Recent L2 studies have predominantly utilized EI tests for two principal objectives: assessing overall language (or oral) proficiency (Bowden, 2016; Davis & Norris, 2021; Gaillard & Tremblay, 2016; McManus & Liu, 2022; Solon et al., 2019, 2022; Tracy-Ventura et al., 2014; Wu & Ortega, 2013; Wu et al., 2023) and measuring implicit knowledge (Ellis, 2005; Erlam, 2006; Spada et al., 2015; Suzuki & DeKeyser, 2015).

Studies validating EI tests as proficiency assessment tools often juxtapose EI scores among distinct proficiency cohorts or examine the relationship between EI scores and other proficiency indicators. Generally, the outcomes of these investigations support the reliability and validity of EI tests. Notably, comprehensive analyses, including those undertaken by Kostromitina and Plonsky (2022) and Yan et al. (2016), amalgamate these findings. Through their meta-analysis of 10 group comparison studies and 11 correlation studies, Yan et al. (2016) observed a consistent trend: learners with higher proficiency consistently outperformed their less proficient counterparts on EI tests and EI scores exhibited a robust correlation with other L2 proficiency indicators. Correspondingly, Kostromitina and Plonsky's (2022) meta-analysis of 46 studies demonstrated a substantial positive correlation between EI tests and alternative proficiency assessments.

Recent research has conducted comparative analyses of EI tests and measures of working memory. Kim et al. (2016) found a noteworthy positive correlation between Korean EI scores and TOPIK speaking and listening scores yet failed to observe a significant link between EI and working memory scores. Park et al. (2020) identified that an oral narrative task yielded better predictions of EI performance than assessments of working memory capacity. They also detected a catalytic effect of short-term memory capacity on less experienced learners, with a relatively diminished impact on more advanced learners. Norris et al. (2023) identified a modest correlation between EI test performance and working memory capacity but a robust correlation between EI performance and the C-test, another proficiency measurement tool. In summary, previous research suggests that while working memory capacity may exert a minor influence on EI test performance, language proficiency significantly influences EI performance.

If Elicited Imitation (EI) tests predominantly measure language abilities rather than working memory, this leads to the question of what specific language skills are being assessed. Wu and Ortega (2013) noted that EI taps into learners' knowledge and their automated use of vocabulary and grammar, delivered with intelligible pronunciation and fluency. They also proposed that the EI test occupies a middle ground between highly communicative tests, like the ACTFL and CEFR proficiency assessments focusing on speaking and listening, and tests that measure more isolated language abilities in non-communicative contexts, such as GJT or C-tests. Ellis (2005), Bowles (2011), and Spada

et al. (2015) demonstrated through factor analysis that implicit grammatical knowledge is the primary target of the EI Test. However, EI tests aimed at measuring implicit knowledge often employ a different format than general proficiency tests, typically concentrating on specific grammatical features and including both grammatical and ungrammatical sentences. Further, Akbary et al. (2023) suggest that EI test outcomes closely resemble those from Oral Paraphrasing tests, which measure listening comprehension. Consequently, while EI test results are closely related to language proficiency, more research is needed to precisely determine the learner abilities they reflect. This emphasizes the importance of a careful interpretation of EI test outcomes.

IRT-based analysis of EI tests

In evaluating learners' linguistic competence through an EI test that employs three or more scoring scales, the polytomous item response model is employed to scrutinize performance outcomes (Campfield, 2017; Deygers, 2020; Graham et al., 2010; Hendrickson et al., 2010; Isbell & Son, 2022; Millard, 2011; Thompson, 2013). IRT analyses were conducted by Graham et al. (2010), Millard (2011), Thompson (2013), and Campfield (2017) using the WINSTEPS program to elucidate the interplay between learner ability and item difficulty as manifested by the EI test. In particular, Millard (2011) and Campfield (2017) evaluated item reliability and difficulty indices to validate the EI test's capacity to distinguish between learners of varying proficiency levels. Given the comparable mathematical rationale underlying the partial credit model (PCM) and the rating scale model (RSM), researchers adopt either of these models according to their research objectives and judgment for result analysis. For example, in contrast to Graham et al. (2008), who employed RSM for EI test analysis, Thompson (2013) opted for PCM-based IRT analysis to assess the distinct roles performed by the scoring scales in the EI test.

The abovementioned studies predominantly concentrated on learner and item parameters, undertaking two-faceted analyses. However, Deygers (2020) and Isbell and Son (2022) employed the FACETS program to explore the influence of rater tendencies on test outcomes. Deygers (2020) managed construct-irrelevant factors by verifying the impact of rater severity and difficulty in scoring criteria on EI test performance using FACETS analysis. Additionally, Isbell and Son (2022) performed a three-faceted IRT analysis grounded in the RSM, encompassing learners, items, and raters as facets. This approach allowed them to appraise learner ability and item difficulty parameters, regulate the rater effect, and identify the factors contributing to score disparities across raters.

Factors influencing EI test performance

Numerous EI investigations indicate a correlation between the length of sentence stimuli or the number of syllables and EI test performance (Campfield, 2017; Davis & Norris, 2021; Graham et al., 2010; Hendrickson et al., 2010; Kim et al., 2016; Ortega et al., 2002; Yan, 2020; Yan et al., 2016). Nevertheless, a recent study by Isbell and Son (2022) provided clarity by asserting that the number of syllables does not serve as a predictive variable for EI performance. Interestingly, Hendrickson et al. (2010) highlighted that numerous linguistic attributes beyond the number of syllables might also have the potential to predict EI test performance. Furthermore, Gaillard and Tremblay (2016)

underscored that the influence of the number of syllables aligns with sentence complexity. Similarly, Graham et al. (2010) viewed sentence length as a moderating factor from a parallel standpoint.

From a linguistic perspective, syntactic attributes of English, encompassing modality, tense, articles, and T-units (Hendrickson et al., 2010), along with factors like the number of clauses and embedded clauses (Isbell & Son, 2022), have been identified as factors that influence EI performance. Additionally, lexical elements, such as lexical frequency (Graham et al., 2010; Hendrickson et al., 2010; Yan, 2020), content or function word density (Campfield, 2017; Graham et al., 2010), and the number of morphemes (Campfield, 2017; Graham et al., 2010; Isbell & Son, 2022), have been investigated as determinants of EI outcomes.

Examining these results from a syntactic vantage point, Hendrickson et al. (2010) discerned the significance of prepositions, followed by aspect and tense, for seven-syllable items. For eight-syllable items, aspect and tense were pivotal, while for nine-syllable items, tense, articles, and the sentence's complexity played significant roles. Conversely, Isbell and Son (2022) did not find that the number of clauses and embedded clauses were significant in explaining item difficulty.

Regarding lexical elements, Graham et al. (2010) and Yan (2020) revealed that lexical frequency could elucidate EI performance in terms of lexical attributes. Nonetheless, the impact of lexical frequency appeared to be influenced by the number of syllables in these studies. Moreover, although Graham et al. (2010) demonstrated that lexical density could elucidate item difficulty, Campfield (2017) did not unearth a significant correlation between content or function word density and difficulty. In the same vein, the influence of the number of morphemes on difficulty was negligible according to Graham et al. (2010), whereas Campfield's study (2017) established a significant correlation between the number of morphemes and content/function word density and item difficulty. However, Isbell and Son (2022) presented a distinct perspective. Alongside vocabulary scores, they identified the number of function morphemes as the most effective variable in explaining item difficulty.

Methodology

Research tools and procedure

The Korean EI test employed in this study is based on the work of Ortega et al. (2002), who formulated congruent EI tests in English, German, Japanese, and Spanish to assess L2 proficiency. Their work showed consistent and robust reliability in EI performance outcomes, effectively discerning discrepancies between lower- and higher-level learner groups across languages. Moreover, Wu and Ortega (2013) and Tracy-Ventura et al. (2014) developed and assessed analogous EI tests in Chinese and French, respectively, and demonstrated the efficacy of these tools in measuring comprehensive L2 oral proficiency.

In a parallel vein, Kim et al. (2016) developed a Korean version of the identical EI test and rigorously validated its reliability and validity. However, owing to certain issues related to English translation, the present study took a proactive stance in refining the EI test. We modified sentence patterns and expressions in the tests used by Ortega et al. (2002) and Kim et al. (2016) to tailor them for the nuances of spoken Korean.

McDade et al. (1982) conducted an inaugural investigation into EI tests and found that participants were able to accurately reproduce even those sentences they had not fully comprehended immediately after hearing them. However, in some cases, participants were asked to replicate sentences after a delay of 3 s. This delay did not significantly influence participants' ability to reiterate sentences they had understood but did adversely impact their capacity to mimic sentences they had not. Recent research indicates that no discernible disparity in EI test performance or perception is attributable to the duration of a delay (Norris et al., 2023). Nevertheless, the standard practice is to structure EI tests with sentences of adequate length to surpass working memory capacity and implement a delay of approximately 2 to 3 s prior to the repetition phase (Park et al., 2020; Yan et al., 2016).

In this study, the Korean EI test consisted of 30 sentences with varying syllable counts, ranging from 8 to 19 syllables. The complete list of sentence stimuli can be found in Additional file 1. These sentences were professionally recorded by a skilled voice actor and played back in sequence, from the 1st to the 30th sentence. To ensure fairness, each sentence was played only once for each participant. After the completion of each sentence, there was a 1-s pause accompanied by a 1-s beep sound. Participants were then given 8 to 10 s to repeat the sentence.[1]

Participants

Learner data were collected from individuals who participated in an evaluation[2] of the EI test devised by this study's research team. To maintain diversity, no more than three learners from the same language background[3] were enrolled, yielding a composition of 15 learners per TOPIK level. Additionally, a cohort of 10 learners was incorporated to assess inter-rater reliability, ensuring that all raters were involved in the scoring process. Table 1 provides an overview of the learner data.

Scoring

Since the Korean EI test is a performance assessment in which learners listen and repeat, the scoring process hinges on raters evaluating learners' speech according to established standards. To ensure the integrity and consistency of the scoring procedure, developing a scoring scale based on well-defined evaluation criteria is imperative.

Traditionally, the evaluation of EI tests has followed a holistic rating paradigm, adopting a rating scale from 0 to 5 points (Lonsdale & Christensen, 2011; McManus & Liu, 2022). When employing an ordinal scale to score EI tests, it is customary to employ three or more scales to evaluate a learner's performance (Chaudron et al., 2005; Graham

[1] A time limit of 10 s was allocated for sentences containing 16 syllables or more.

[2] The EI test devised by the researchers was employed as a component of a speaking education evaluation that aimed to gather "Korean voice data of Western and Asian language users for language education." This initiative was part of a support project overseen by the National Information Society Agency (NIA) to collect artificial intelligence learning data. A comprehensive collection of 3000 h of voice data was created, along with accompanying metadata detailing learners' nationality, mother tongue, and TOPIK level. The NIA provided explanations of the study to all participants and obtained their consent to participate.

[3] The learner cohort had a diverse range of mother tongues, including but not limited to German, Russian, Lithuanian, Malay, Mongolian, Burmese, Bulgarian, Vietnamese, Spanish, Slovak, Armenian, English, Estonian, Ukrainian, Italian, Indonesian, Japanese, Chinese, Czech, Kazakh, Khmer, Telugu, Turkish, Persian, Portuguese, Polish, French, Hungarian, and Hindi.

Table 1 Learner information

Group	Proficiency (TOPIK level)	*N*	Total
Beginner	1	15	30
	2	15	
Intermediate	3	15	30
	4	15	
Advanced	5	15	30
	6	15	
Common scoring data		10	10
			100

et al., 2008). Several studies on EI tests have used the test developed by Ortega et al. (2002) (Bowden, 2016; Kim et al., 2016; McManus & Liu, 2022; Park et al., 2020; Tracy-Ventura et al., 2014; Wu & Ortega, 2013). In their experimental procedures and subsequent analyses, these studies employed a scale ranging from 0 to 4 points.

In this study, we adopted an analytical scoring approach instead of the conventional holistic scoring method for EI tests. This involved a breakdown of constructs into "content" and "delivery." Just as content elements like vocabulary and grammar are pivotal for assessing the extent of native speech imitation in EI tests, delivery components such as intonation, stress, and segmental pronunciation are also significant criteria. Therefore, we contend that through analytically examining these criteria, the degree of restoration can be effectively validated (Burger & Chretien, 2001; Gaillard & Tremblay, 2016). Based on this framework, we established the delivery construct as a scoring criterion for evaluating the degree of restoration within EI tests, employing an analytical scoring methodology.

To mitigate any simplistic, impression-based evaluations, we endeavored to articulate scoring criteria as explicitly and quantifiably as feasible. We disentangled the content and delivery constructs, establishing rubrics on a 0–5-point scale that corresponded with the level of restoration in terms of meaning alterations.[4] We supplemented this with concrete scoring instances across diverse scoring scenarios. Within the delivery construct, a learner's restoration was most successful when their speech closely mirrored that of a native speaker. To facilitate the nuanced evaluation of advanced learners, we introduced a highest score of 5 points, thereby adopting a broader scale than the 0–4-point scale utilized by Ortega et al. (2002).

Ahead of the actual scoring, the raters underwent three rounds of rigorous training. The initial phase involved comprehensive workshops that explained the precise scoring criteria for content and delivery, which were supplemented by illustrative examples. In

[4] Scores for the content construct were allocated as follows: 5 points if the original content was fully restored, 4 points for lexical or grammatical errors that did not alter the core meaning, 3 points if the core meaning changed but over 50% of the original intent was recovered, 2 points if less than 50% was restored and the primary meaning was barely conveyed, 1 point for only minimal restoration where the original meaning remained unexpressed, and 0 points if evaluation was not possible. For the delivery construct, the evaluation criteria encompassed elements such as pronunciation, intonation, speech rate, pauses, and syllable segmentation. Scores were distributed as follows: 5 points for pronunciation with native-like fluency, 4 points if delivery diverged from native delivery but contextual inference was unnecessary to understand the restored sentence, 3 points if some context-based inference was necessary due to delivery, 2 points if context-based inference was challenging and delivery appeared very unnatural, 1 point if context-based inference was unattainable, and 0 points if evaluation was not possible.

the subsequent round, raters reviewed specific examples of scoring in Korean EI tests to fully acquaint themselves with the scoring criteria. The final stage entailed thorough training on the precise scoring process and the structural dynamics of forthcoming scoring data. This final round of training enabled the raters to develop a comprehensive grasp of the study's overarching content related to EI tests and seamlessly execute the scoring process.

A total of 100 learners were evaluated by 22 raters using a partial crossover method. To appropriately account for rater influence in IRT analyses, establishing a consistent framework in which raters share identical scoring practices is crucial (Lee, 2008). In line with this rationale, all 22 raters collectively scored 10 learners. For the remaining 90 learners, we structured the scoring data so that two to three raters could consistently assess each learner within designated groups.

Analysis

Correlation analysis and analysis of variance

To address research question 1 vis-à-vis the TOPIK, we examined the correlation between the mean score of the EI test and the TOPIK proficiency level (ranging from 1 to 6). Additionally, we sought to elucidate disparities in mean EI test scores across beginner, intermediate, and advanced levels. The TOPIK assessment is segmented into TOPIK I (for beginners) and TOPIK II (for intermediate and advanced learners). In total, the TOPIK consists of 6 levels; TOPIK I comprises levels 1 to 2, and TOPKI II comprises levels 3 to 6. It is worth noting that TOPIK I evaluates listening and reading, while writing, listening, and reading are all appraised in TOPIK II. Similar to our study, Kim et al. (2016) used the TOPIK listening score due to the auditory nature of EI tests. However, our investigation adopted the comprehensive final level derived from evaluations across all TOPIK sections (listening, reading, and writing) to holistically assess overall language competence. Cases without TOPIK scores were treated as missing data and subsequently excluded from the correlation analysis. Considering the use of analytical scales that include content and delivery construct scores, we computed the EI test's mean score by averaging these two categories. To evaluate discrepancies among the three groups, we performed a one-way analysis of variance (ANOVA), setting the confidence level at 95% and the significance level (α) at 0.05.

Many facets rasch model analysis

To comprehensively explore the inherent characteristics and patterns within each facet (learner, item, rater, and construct) of the Korean EI test and address research question 2, we used the Many Facets Rasch Model (MFRM) for analysis. In order to employ the MFRM, it is imperative to satisfy two assumptions postulated within the Rasch model framework(Eckes, 2015:27). This study posits that the Korean EI test serves as a tool for assessing proficiency and accordingly asserts the fulfillment of the first assumption, namely the unidimensionality assumption. Furthermore, given the observed independence among the items comprising the Korean EI test, it is conceivable to affirm the satisfaction of the second assumption, namely the local independence assumption. The MFRM constitutes a one-parameter IRT model with an emphasis on difficulty. Notably, item discrimination is fixed at 1 within this model. The MFRM operates on a

probabilistic foundation, enabling the estimation of latent competence—essentially a latent variable—by leveraging difficulty parameters. This approach facilitates the exploration of parameter values across diverse facets within a controlled framework. Since this study necessitated the simultaneous analysis of multiple facets, a characteristic trait of MFRM analysis, adopting the MFRM aligns with the broader research landscape (Deygers, 2020; Isbell & Son, 2022; Solon et al., 2022). Our analysis used the FACETS program (version 3.80.4; Linacre, 2016), which enabled an in-depth exploration of the data related to each of the four facets of the Korean EI test.

Given that the Korean EI test formulated for this study involved a polytomous response model, scored on a 0–5 scale, MFRM analysis was considered appropriate. In this analytical approach, logit scores are ascertained based on the probabilistic framework, providing a basis for analyzing parameter values within each facet. Following Graham et al.'s (2008) rationale, which suggested the appropriateness of the RSM for examining EI test items, our analysis centered on an RSM-based framework, addressing the four key facets of the Korean EI test. Operating within a probabilistic domain, RSM accommodates the prospective scores attainable across the scoring scale in conjunction with item difficulty. Therefore, our analysis involved a thorough investigation into learners' ability scores, item difficulty, and rater severity and reliability, alongside factors germane to the rater facet, such as rater separation reliability.

Explanatory item response modeling approach

To investigate the factors influencing EI test performance as outlined in research question 3, we conducted a detailed analysis of linguistic factors. Before this analysis, we delineated and curated linguistic factor definitions, a crucial preparatory step. First, we considered the number of syllables, a factor commonly cited in prior studies (Campfield, 2017; Davis & Norris, 2021; Graham et al., 2010; Hendrickson et al., 2010; Kim et al., 2016; Ortega et al., 2002; Yan, 2020; Yan et al., 2016). Following Campfield's (2017) assertion that the number of words more effectively encapsulates sentence length than syllable count, we adapted this metric to the Korean language by integrating the number of "ecel" (a space-based word unit in Korean) in lieu of words. Regarding syntactic dimensions, we included the number of clauses and embedded clauses (Isbell & Son, 2022) for lexical aspects, we included the number of morphemes as a salient influencing factor, following the observations of many researchers (Campfield, 2017; Graham et al., 2010; Isbell & Son, 2022).

To quantify lexical frequency (Graham et al., 2010; Hendrickson et al., 2010; Yan, 2020), we counted the number of words aligned with each level of the Korean vocabulary list for learners (NIKL, 2003). This count was then subjected to level-specific weights (levels 1–2: 1 point, levels 3–4: 2 points, levels 5–6: 3 points) to calculate the vocabulary score, following the framework described by Isbell and Son (2022). In parallel, we evaluated grammatical elements by assigning level-associated weights, as outlined in the International Standard Curriculum of Korean Language (National Institute of the Korean Language, 2017). These constructs comprised several components: vocabulary encompassed nouns, pronouns, numerals, verbs, adjectives, determiners, adverbs, exclamations, and affixes, while grammar encompassed particles, endings, and expressions.

Notably, when vocabulary was incorporated within an expression, it was not treated as an independent vocabulary entity.

To consider the influence of various factors, our approach also drew on insights from previous research (Campfield, 2017; Graham et al., 2010; Isbell & Son, 2022), including the total number of morphemes, content words, function words, and the density of content words and function words. For uniformity and coherence, we delineated content words and function words according to the guidelines of the Korean vocabulary list for learners (NIKL, 2003) and the International Standard Curriculum of Korean Language (NIKL, 2017)[5]

For the statistical analysis, we employed the EIRM package within the R programming environment (Bulut, 2021; Bulut et al., 2021). This specialized package affords regression analysis functionality complemented by the RSM, a quintessential component of IRT analysis. Notably, this approach facilitates the examination of influential factors through multiple linear rating scale models (LRSMs). It also overcomes the limitations tied to diminished power and precision that arise when transforming numerous individual item responses into a confined n-size of 30, which corresponds to the number of items (Isbell & Son, 2022, p. 876).

The initial step in our statistical analysis involved a correlation analysis of the identified influencing factors. Subsequently, the first LRSM was constructed with reference to Isbell & Son (2022), incorporating parameters such as the number of syllables, vocabulary score, embedded clauses, the number of content and function words, and the number of clauses—each of which was recognized as a factor impacting the observed EI score. The EI score was calculated by averaging the scores attributed to the content and delivery constructs. During the analysis, subsequent factors such as the number of ecels, grammar score, the density of content and function words, and the number of morphemes were either introduced or excluded based on their statistical significance or the estimated impact. The primary aim was to ascertain the most optimal model, characterized by the highest correlation between the difficulty parameters anticipated through the EI test and the descriptive difficulty parameters emanating from the model. Throughout the analysis, a confidence level of 95% and a significance level (α) of 0.05 were maintained, ensuring rigor and consistency in the interpretation of results.

Results

Prior to the analysis, a crucial preliminary step involved confirming the internal reliability of the newly formulated EI test. The Cronbach's α values for the EI tests demonstrated robust internal consistency. Specifically, for the content and delivery constructs, Cronbach's α was 0.981 and 0.986, respectively; thus, both were comfortably above the recommended threshold of 0.7.

[5] In this study, eight morphemes outside of the NIKL's classification (2017) (e.g., causative affix "-i-", plural affix "tul", dependent noun "li", etc.) were included in the number of content/function words and morphemes but were excluded from the vocabulary and grammar scores.

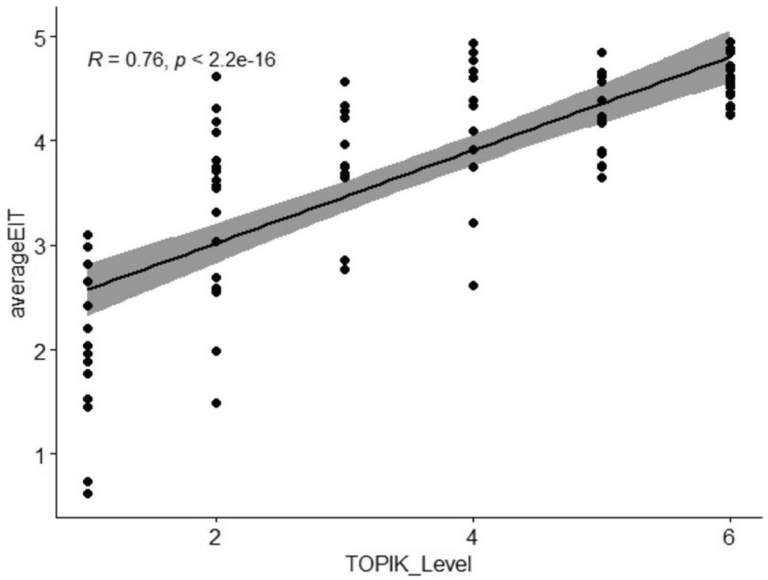

Fig. 1 Correlation between TOPIK levels and average EI scores

Results of correlation analysis and ANOVA

To elucidate the efficacy of the EI test as a proficiency measurement tool, we investigated the relationship between TOPIK levels and the mean scores of the test. Since TOPIK levels represent ordinal data, we conducted a Spearman rank correlation analysis. We chose this type of analysis because it had the capacity to circumvent distribution assumptions while probing the linkage between TOPIK levels and the mean scores of the EI test.

The analysis revealed a robust and statistically significant correlation of 0.76 between TOPIK levels and the mean scores of the EI test ($p = 0.000 < 0.05$). Notably, this correlation underscored an alignment between TOPIK proficiency levels and the performance reflected in the EI test's average scores. The scatterplot in Fig. 1 visually depicts this relationship.

We then examined whether the average EI scores differed according to TOPIK level (beginner, intermediate, and advanced). Table 2 and Fig. 2 present the descriptive statistics for the average EI scores by TOPIK level.

The results of a one-way ANOVA revealed a significant difference between the beginner, intermediate, and advanced groups ($F_{(2,91)} = 46.41$, $p = 0.000 < 0.05$, $\eta^2 = 0.51$). The Tukey post hoc test also showed significant differences between the beginner and intermediate groups ($p = 0.000 < 0.05$), as well as between the beginner and advanced groups ($p = 0.000 < 0.05$). However, the difference between the intermediate and advanced groups was not significant ($p = 0.09 > 0.05$). Table 3 presents the post-hoc test results.

IRT analysis results

Utilizing the MFRM-based FACETS program, we obtained valuable logit information concerning the study participants, raters, items, and constructs, as depicted in Fig. 3. Distribution was strategic so that advanced learners, challenging items, stringent raters, and demanding constructs were positioned at the upper end of the spectrum.

Table 2 Descriptive statistics of average EI scores by TOPIK level

Group (TOPIK level)	*N*	M	SD	SE
1 (Beginner: 1–2)	32	2.7	1.05	0.19
2 (Intermediate: 3–4)	28	4.01	0.63	0.12
3 (Advanced: 5–6)	32	4.41	0.34	0.06

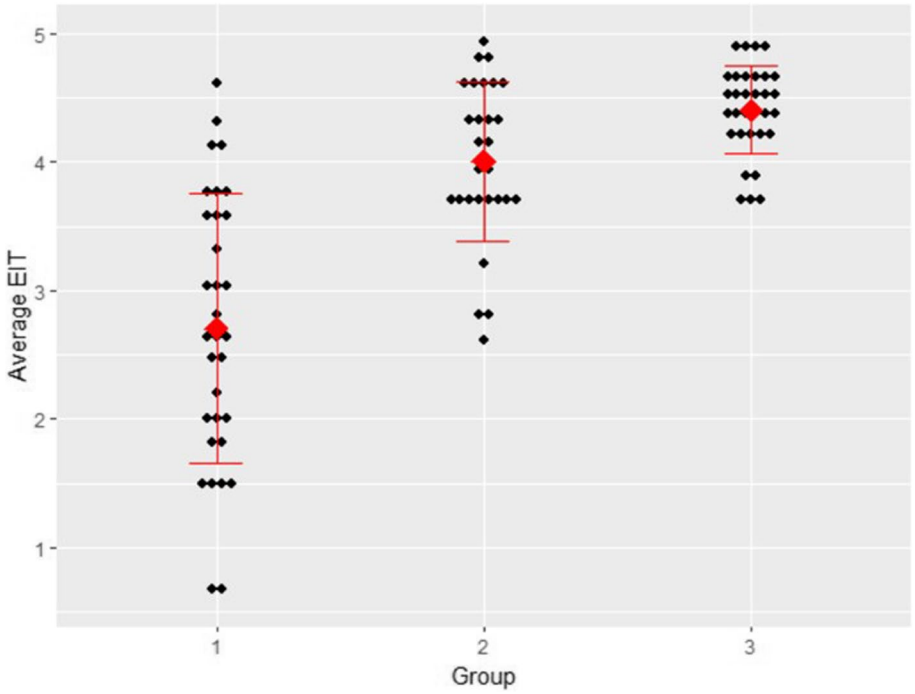

Fig. 2 Scatter plot of average EI scores by TOPIK level

Table 3 Tukey post hoc test results by group

	b	SE	95% CI	*t*	*P*
2–1	1.30	0.19	0.84—1.76	6.83	***
3–1	1.70	0.18	1.26—2.14	9.23	***
3–2	0.40	0.19	-.05—0.86	2.10	0.09

In the second column of the learner facet, a large proportion of learners fall within the 1 to 2 logits range. Notably, a thorough examination of the precise learner parameter estimates calculated by the FACETS program showed that all learners, except for eight individuals, displayed values exceeding 0 logits.

In the third column, the set of 30 items employed in this study displays an equitable distribution spanning roughly from −1 to 0.5 logits. This distribution indicates that the items were optimally suited for learners at their corresponding ability levels. The application of IRT analysis facilitated the direct comparison of item difficulty and learner ability on a shared scale. Notably, the analysis revealed that the most challenging item had a difficulty rating of 0.51 logits, subsequently showing that 87 learners possessed

Fig. 3 Facet elements measurement

Table 4 IRT-based item difficulty

Item	I01	I02	I03	I04	I05	I06	I07	I08	I09	I10
Difficulty	−0.68	−1.07	−0.64	−0.63	−0.37	0.12	0.07	−0.86	0.02	−0.49
Item	I11	I12	I13	I14	I15	I16	I17	I18	I19	I20
Difficulty	0.27	−0.11	0.23	0.32	0.35	0.16	−0.51	−0.01	0.39	0.36
Item	I21	I22	I23	I24	I25	I26	I27	I28	I29	I30
Difficulty	0.37	0.46	0.45	0.22	0.26	0.20	0.19	0.32	0.51	0.13

latent abilities surpassing this level. This outcome underscores that the participating learners exhibited a commendable performance in the EI tests, further affirming their overall proficiency.

Due to the absence of predefined answers for each score scale in the EI test, the estimation of a learner's ability can vary depending on different raters. To assess the influence of rater severity on EI performance, we conducted an analysis of the scoring behaviors exhibited by the 22 raters. The results confirmed that these raters generally clustered around the 0 logits point. In the distribution illustrated in Fig. 3, a positioning at 0 logits signifies that the rater's impact is minimal. Consequently, the individual inclinations of the raters had a limited effect on the evaluation of learners' EI test performance.

In the final evaluation, the raters assessed learners' EI test performances based on both content and delivery constructs. The "Criteria" section of the fifth column shows that the delivery construct received slightly more stringent scoring than the content construct.

Table 4 shows the findings of the analysis of item difficulty parameters. Using RSM, the analysis revealed that the distribution of item difficulty spanned from the difficulty measure of item 2 ["책이 책상 위에 있다. (chayki chayksang wiey issta: the book is on the table)"], which was –1.07 logits, to the difficulty value of item 29 ["열한 시 반 기차가 이미 역을 떠났는지 모르겠다. (yelhan si pan kichaka imi yekul ttenassnunci molukeyssta: I don't know if the 11:30 train has left the station yet)"], at 0.51 logits.

Figure 4 illustrates comparisons between the difficulty indices of individual items and their corresponding observed averages. This study employed a scoring range of 0 to 5; therefore, the highest attainable average score per item was 5 points.[6] Our analysis confirmed that with the exception of items 22 (observed average: 3.48), 23 (observed average: 3.49), and 29 (observed average: 3.41), all other items had averages of 3.5 or higher. These findings elucidate the underlying reason for the relatively low difficulty indices associated with the items.

Table 5 presents an overview of the evaluation tendencies demonstrated by the participating raters. The mean values, adjusted by applying MFRM, ranged from 3.95 to 4.43. Notably, when using the 0–5-point scale, the divergence between the most stringent and most lenient raters was approximately 0.5. The estimates to rater parameters fell within the range of 0.28 to –0.29. This distribution implies that rater severity exerted an influence on learners' scores, which, in turn, followed a standard normal distribution, varying by approximately ± 0.3 standard deviations.

[6] Scoring in this study used a 0–5 scale for both content and delivery constructs. However, the final score was calculated by averaging the sum of scores for both two constructs, yielding a maximum score of 5 for learners.

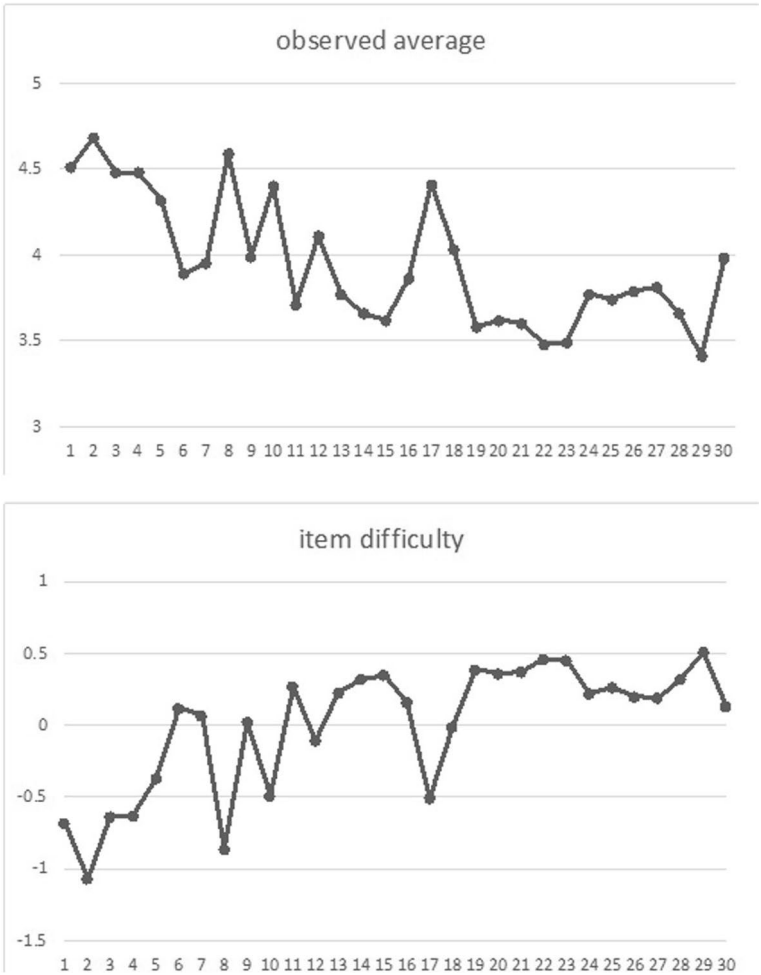

Fig. 4 Difficulty indices of individual items and their corresponding observed averages

The population separation ratio of 5.15 indicates that the variations in severity among raters were roughly five times more significant than the measurement error. Moreover, the separation reliability of 0.96, and the strata value of 7.20 demonstrate that the raters had distinctive evaluation patterns. These analysis outcomes suggest that the raters in this study approached scoring with consideration to their individual severities. Given that the MFRM is employed to mitigate rater stringency/leniency effects on learner scores (Lawson & Brailovsky, 2006:651), it can be affirmed that the reliability of the raters participating in this study has been ensured. Consequently, it can be ascertained that there is no deviation from fairness in the learners' ultimate fair score.

The ideal infit mean square value, which serves as a significant indicator of rater reliability, is between 0.5 and 1.5. This range signifies that raters have executed assessments with robust internal consistency (Kondo-Brown, 2002; Lunz & Stahl, 1990; Weigle, 1998;). Our analysis confirmed that all raters in this study met this criterion, thus ensuring a high level of internal consistency among them.

Exploring the results of the FACETS program analysis showed that the actual and anticipated agreements among raters were 40.3% and 33.6%, respectively. Consequently,

Table 5 IRT-based rater parameters

Rater	Observed average	Fair (M) average	Measure	SE	Infit MnSq
R05	3.98	3.95	0.28	0.03	1.08
R19	3.46	3.96	0.27	0.03	1.42
R09	3.93	3.99	0.24	0.03	1.19
R16	3.70	4.04	0.19	0.03	1.26
R03	3.80	4.07	0.17	0.03	1.21
R11	3.80	4.07	0.16	0.03	1.01
R07	3.94	4.10	0.13	0.03	1.14
R14	4.08	4.17	0.06	0.02	0.95
R21	3.64	4.20	0.02	0.03	1.39
R17	3.75	4.21	0.01	0.04	1.38
R10	3.79	4.21	0.01	0.03	0.91
R08	4.10	4.22	0.00	0.03	0.80
R13	3.80	4.25	−0.03	0.03	0.98
R12	3.96	4.27	−0.06	0.03	0.72
R01	4.16	4.29	−0.09	0.03	0.93
R20	3.85	4.32	−0.12	0.04	0.84
R02	4.12	4.35	−0.16	0.02	0.85
R04	4.13	4.36	−0.17	0.03	0.83
R22	3.94	4.36	−0.18	0.04	1.09
R15	4.06	4.38	−0.21	0.03	0.86
R06	4.18	4.40	−0.23	0.03	0.89
R18	3.86	4.43	−0.29	0.04	0.95

the real agreement among raters surpassed the expected level of agreement in the Rasch model. When converted into logits, the inter-rater agreement equated to a logit value of 0.100, suggesting high reliability both within and among the raters in this study. Consequently, we concluded that in terms of severity levels, the participating raters appropriately discerned learners' variations while upholding rater reliability.

Table 6 shows the results of the analysis of patterns in learners' EI test performance, categorized into content and delivery constructs. The adjusted mean values for the delivery and content constructs were 4.09 and 4.33, respectively, indicating that delivery was evaluated more rigorously than content. Nevertheless, upon being subjected to adjustments via the Rasch model, both constructs exceeded an average value of 4 and displayed no noteworthy disparity from the overall average.

Scrutinizing these outcomes through the lens of the raters' scale utilization behaviors showed that scores of 4 and 5 contributed to approximately 70% of the total scale utilization. This finding offers insight into the elevated average scores for the constructs, the commendable performance of the learners, and the lenient evaluation tendencies exhibited by the raters.

Results of EIRM analysis

Table 7 presents the results of the correlation analysis examining the linguistic factors that impacted EI test performance. The number of syllables had significant correlations

Table 6 IRT-based item statistics

Construct	Observed Average	Fair (M) Average	Estimate	SE
Delivery	3.81	4.09	0.14	0.01
Content	4.09	4.33	−0.14	0.01

Table 7 Correlation analysis of linguistic factors influencing EI test performance

	1	2	3	4	5	6	7	8	9	10	11
1	-										
2	.68**	-									
3	.38*	.47**	-								
4	.24	.32	.84**	-							
5	.85**	.86**	.47*	.37*	-						
6	.75**	.68**	.16	.13	.68**	-					
7	.76**	.50**	.58**	.48**	.78**	.40*	-				
8	.66**	.92**	.31	.23	.85**	.69**	.41*	-			
9	.79**	.56**	.50**	.40*	.87**	.48**	.90**	.48**	-		
10	−.31**	.15	−.21*	−.17	−.22*	−.00	−.62**	.31	−.64**	-	
11	.31**	−.15	.21*	.17	.22*	.00	.62**	−.31	.64**	−1**	-

1 number of syllables, *2* number of ecels, *3* number of clauses, *4* number of embedded clauses, *5* number of morphemes, *6* vocabulary score, *7* grammar score, *8* number of content words, *9* number of function words, *10* content word density, *11* function word density

* $p < .05$, ** $p < .01$, *** $p < .001$

with all the factors, except for the number of embedded clauses. Moreover, the number of morphemes was significantly correlated with all the other factors.

Five LRSMs were constructed following the principles of EIRM. In LRSM 1, several factors from Isbell and Son (2022) were considered, including the number of syllables, vocabulary score, embedded clauses, content words, function words, and clauses. Subsequently, factors with non-significant p-values—clauses ($p = 0.95 > 0.05$), embedded clauses ($p = 0.77 > 0.05$), content words ($p = 0.35 > 0.05$), and function words ($p = 0.11 > 0.05$)—were removed.

In LRSM 2, the analysis incorporated the number of ecels and grammar score. However, in LRSM 3, the number of ecels ($p = 0.06 > 0.05$) and grammar score ($p = 0.10 > 0.05$) were excluded due to their lack of significance, and content word density and function word density were introduced instead. Content word density displayed significance ($p = 0.000 < 0.05$), while function word density did not ($p = 0.99 > 0.05$). Consequently, LRSM 4 discarded function word density and included the number of morphemes. However, the number of morphemes failed to achieve significance ($p = 0.74 > 0.05$) and was excluded from LRSM 5.

The highest explanatory power emerged in LRSM 2 ($r = 0.892$); however, this model contained two non-significant factors. The explanatory power remained consistent from LRSM 3 (which incorporated content word density) to LRSM 5 ($r = 0.887$). As such, LRSM 3, which had the greatest number of significant factors among all the models, was identified as the optimal model. Table 8 provides a comprehensive overview of the analysis results across the five LRSMs.

Regarding the significant factors identified in LRSM 3, as the number of syllables and vocabulary score increased, the complexity of items also rose, with the vocabulary score exerting a more pronounced impact on EI performance than the number of syllables. A reduction in content word density was associated with an increase in the difficulty of items. Figure 5 visually illustrates the correlation between the predicted difficulty parameters of the EI test and the descriptive difficulty parameters elucidated by LRSM 3.

Discussion

Correlation between EI test scores and TOPIK levels and patterns of EI test performance by group

This study aimed to assess the validity of a developed EI test as an indicator of language proficiency by investigating its association with TOPIK levels and analyzing the patterns of test performance across different learner groups. The research question 1 focused on establishing a correlation between TOPIK levels, derived from learners' metadata, and the mean scores achieved in the EI test. We found a strong positive correlation of 0.76 between TOPIK levels and the mean EI scores ($p = 0.000 < 0.05$), underscoring the developed EI test's potential as a reliable measure of language proficiency.

It is important to acknowledge the diverse range of proficiency measurement indicators used in previous studies, such as spoken narrative test scores (measured using complexity, accuracy and fluency, or speech rate) (McManus & Liu, 2022; Park et al., 2020; Tracy-Ventura et al., 2014; Wu & Ortega, 2013; Wu et al., 2022), adjusted standardized test scores (Davis & Norris, 2021; Kim et al., 2016; Solon et al., 2019, 2022), C-test scores (Davis & Norris, 2021; Gaillard & Tremblay, 2016; Norris et al., 2023), interview scores, Read Aloud scores, other speaking test scores (Bowden, 2016; Davis & Norris, 2021; Tracy-Ventura et al., 2014), writing test scores (Tracy-Ventura et al., 2014), and final grades (Tracy-Ventura et al., 2014). All of these indicators have previously been correlated with EI scores. However, our study narrows its focus to the relationship between EI tests and learners' existing proficiency levels, omitting simultaneous measurement tests. Despite this limitation, the robust correlation between TOPIK levels and EI test scores reinforces the credibility of our EI test as a valid tool for assessing language proficiency.

We further explored the distinguishing patterns of EI test performance among different learner groups. We conducted a one-way ANOVA, grouping learners into three categories: beginners (TOPIK levels 1–2), intermediates (TOPIK levels 3–4), and advanced (TOPIK levels 5–6). Notably, we found significant differences between beginner and intermediate learners, as well as between beginner and advanced learners. However, no significant distinction emerged between intermediate and advanced learners. This suggests that although the EI test effectively discerned proficiency gaps between beginners and intermediates, its discriminatory effectiveness waned when comparing learners with higher proficiencies.

These findings diverge slightly from the results of a meta-analysis conducted by Yan et al. (2016), which indicated that among 10 studies comparing EI test performance across learner groups, most used EI test tasks to differentiate between advanced and beginner or intermediate groups. Only two studies (Iwashita, 2006; Serafini, 2013) employed EI tests to distinguish between beginner and intermediate learners. The current study's findings support the validity of the developed EI test as a proficiency

Table 8 EIRM results for EI test item difficulty

	LRSM 1			LRSM 2			LRSM 3			LRSM 4			LRSM 5		
	Est	**SE**	***p***	**Est**	**SE**	***p***	**Est**	**SE**	***p***	**Est**	**SE**	***p***	**Est**	**SE**	***p***
1	0.06	0.03	*	0.06	0.03	*	0.06	0.02	***	0.06	0.03	*	0.06	0.02	**
2	0.12	0.04	**	0.13	0.04	***	0.11	0.04	***	0.11	0.04	**	0.12	0.04	**
3	− 0.01	0.10	0.95												
4	− 0.03	0.10	0.77												
5	− 0.04	0.04	0.35												
6	0.08	0.05	0.11												
7				− 0.08	0.05	0.06									
8				0.06	0.04	0.10									
9							− 1.17	0.32	***	− 1.21	0.62	0.05	− 1.21	0.62	0.05
10							0.01	0.44	0.99						
11										0.01	0.03	0.74			
0/0.5	14.33	829.91	0.99	14.42	817.00	0.99	14.97	815.84	0.99	14.98	814.60	0.99	14.98	817.56	0.99
0.5/1	− 21.22	465.84	0.96	− 21.13	473.54	0.97	− 20.57	475.45	0.97	− 20.56	475.13	0.97	− 20.56	473.03	0.97
1/1.5	− 2.26	0.33	***	− 2.15	0.32	***	− 1.58	0.21	***	− 1.58	0.49	**	− 1.58	0.49	***
1.5/2	− 4.04	0.32	***	− 3.91	0.31	***	− 3.34	0.20	***	− 3.34	0.48	***	− 3.35	0.48	***
2/2.5	− 2.59	0.30	***	− 2.48	0.29	***	− 1.91	0.17	***	− 1.91	0.47	***	− 1.91	0.47	***
2.5/3	− 3.37	0.29	***	− 3.26	0.28	***	− 2.69	0.16	***	− 2.68	0.46	***	− 2.69	0.47	***
3/3.5	− 2.51	0.27	***	− 2.40	0.26	***	− 1.83	0.13	***	− 1.82	0.45	***	− 1.82	0.45	***
3.5/4	− 2.62	0.26	***	− 2.51	0.25	***	− 1.95	0.11	***	− 1.94	0.45	***	− 1.94	0.45	***
4/4.5	− 1.84	0.25	***	− 1.73	0.24	***	− 1.17	0.09	***	− 1.16	0.45	**	− 1.16	0.45	**
4.5/5	− 0.66	0.24	**	− 0.55	0.23	*				0.02	0.44	0.96	0.02	0.45	0.96
r	.886		***	.892		***	.887		***	.887		***	.887		***

1 number of syllables, *2* vocabulary score, *3* number of clauses, *4* number of embedded clauses, *5* number of content words, *6* number of function words, *7* number of ecels, *8* grammar score, *9* content word density, *10* function word density, *11* number of morphemes, *r* correlation coefficient *$p < .05$, **$p < .01$, ***$p < .001$

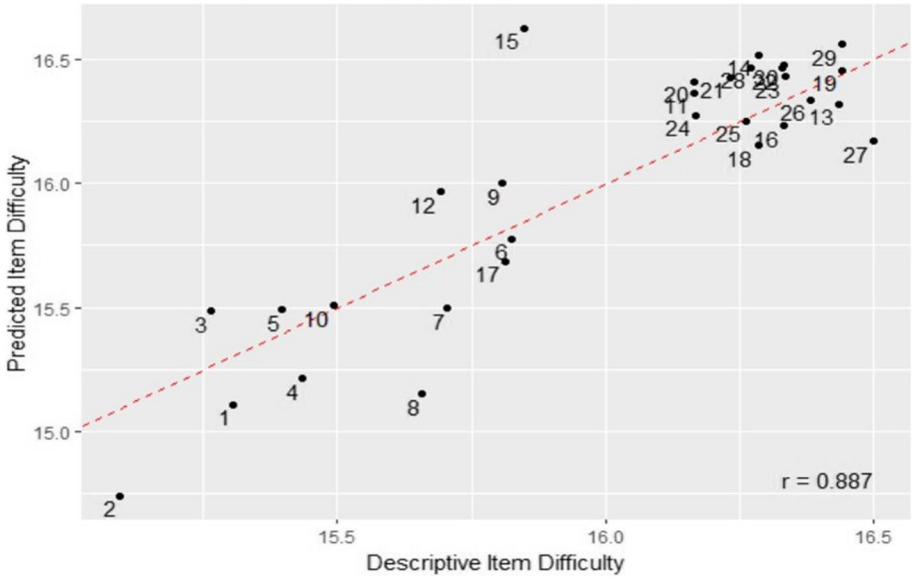

Fig. 5 Scatter plot of descriptive and predictive item difficulties

measurement tool that can distinguish beginner and intermediate learners. However, as noted by Yan et al. (2016), caution must be exercised when generalizing the test's discriminatory ability to all proficiency levels. Solon et al. (2019) examined a different proficiency range by including advanced and higher language proficiencies from earlier studies and revealed that the EI test modified by Bowden (2016) from Ortega et al. (2002) might not be suitable to discern nuanced advanced-level proficiency differences. Solon et al. (2019) suggested increasing the length and difficulty of EI items to improve discriminatory ability at higher proficiency levels. To this end, we propose that the Korean EI test developed in our study, drawing on the insights of Ortega et al. (2002) and Kim et al. (2016), should incorporate more intricate and extensive items to effectively distinguish among learner groups with proficiencies beyond the intermediate level.

Facet-specific patterns of the EI test according to IRT

The research question 2 describes the facet-specific dynamics of the Korean EI test investigated in this study. When assessing item difficulty, a range of − 1.07 to 0.51 logits emerged. Considering that a substantial proportion of learners exhibited abilities surpassing 0.51 logits, the selected items may not have effectively discriminated among the abilities of these proficient learners. It is worth clarifying that item difficulty denotes the ability level at which an item's functionality is optimized (Baker, 2001). In this context, the EI test items generally functioned suitably to assess learners ranging between − 1.07 and 0.51 logits. However, to accurately measure the potential capacities of learners with higher proficiency, the incorporation of more challenging questions is warranted.

Nevertheless, as shown in Fig. 3, the difficulty of more demanding items did not surpass that of less challenging items. This finding underscores the need to scrutinize the factors influencing the difficulty of this study's EI test and enhance the difficulty of items positioned above the 0-logit threshold.

In a related study, Isbell and Son (2022, p. 873) investigated the EI test created by Kim et al. (2016) using IRT analysis. They found that items 13, 17, 19, 22, 15, and 16, positioned between 1 and 2 logits, exhibited higher item difficulty than those in the present study, as evidenced by their distribution. Interestingly, item 17 emerged as one of the most challenging items in Isbell and Son's research (2022) yet was among the least challenging items in our study, with a difficulty value of −0.51. This disparity can be attributed to the sentence stimuli "내가 지금 만나는 사람은 재미가 있다. (nayka cikum mannanun salamun caymika issta: The person I'm dating has a wonderful sense of humor, 15 syllables)" used in our study versus "내가 현재 만나고 있는 분은 재치가 있다. (nayka hyencay mannako issnun pwunun caychika issta: The person I'm dating has a wonderful sense of humor, 16 syllables)" used in Kim et al.'s (2016) study. The distinct difficulty patterns of these two sentences, despite their similar meanings and syllable counts, emphasize the influence of vocabulary and grammar on EI test performance. Evidently, the process of adapting sentences into natural spoken language in our study likely contributed to the reduction in item difficulty. This finding underscores the importance of considering linguistic factors such as vocabulary and grammar when translating EI tests across languages or devising equivalent EI tests, as expounded upon later in this discussion.

In terms of the rater facet, we observed that all raters effectively differentiated learners based on severity level, ensuring both intra-rater and inter-rater reliabilities at a notable level. This reaffirms the feasibility of establishing reliable EI test scoring on binary and ordinal scales. Such streamlined scoring procedures can enhance the practical utility of EI tests within both classroom and research settings.

In contrast to previous studies, this investigation employed analytical scoring of the EI test using two constructs—content and delivery. The findings revealed a slightly stricter scoring approach for delivery than for content. Nevertheless, we observed no significant difference between these two constructs, given that the average scores, adjusted within the Rasch model, surpassed 4 points for both constructs. This outcome can be attributed to raters frequently using scales of 4 and 5 points in their comprehensive scoring. It also reflects the learners' exceptional performance and the raters' lenient evaluation tendencies. As such, this study could not definitively conclude whether analytical scoring exhibits different tendencies in comparison to holistic scoring. For further insights, future investigations should explore how the scoring of content and delivery constructs functions when learners take EI tests of elevated difficulty beyond their current proficiency level.

A limitation of this study, thus, lies in the reduced item difficulties incorporated during the modification of the Korean EI test. In the future, the development of an EI test tailored to advanced learners and the incorporation of sophisticated cross-scoring methodologies should be used to further validate the EI test.

Linguistic factors influencing EI test performance

The research question 3 focuses on the linguistic factors that contribute to EI test performance. Prior studies have explored the influence of different linguistic elements, yielding mixed results. Syntactic factors, such as the number of syllables, have demonstrated varying effects: Hendrickson et al. (2010) reported an impact,

while Isbell and Son (2022) found that these factors had no significant influence. Among lexical factors, lexical frequency has emerged as a pivotal factor affecting test difficulty (Graham et al., 2010; Yan, 2020). Notably, studies have presented contrasting findings concerning the impact of density, with both significant (Graham et al., 2010) and insignificant (Campfield, 2017) effects reported. Similarly, the role of the number of morphemes has differed among studies, being influential in some (Campfield, 2017) and inconsequential in others (Graham et al., 2010). Studies have also explored the impact of the number of content and function words on test performance (Campfield, 2017; Isbell & Son, 2022), and one study specifically assessed the influence of vocabulary scores (Isbell & Son, 2022).

In this study, we conducted a comprehensive correlation analysis of all variables that displayed discrepancies in earlier research as potential linguistic factors. Adopting the EIRM approach, we constructed multiple LRSMs to identify the model that best elucidated the difficulty of EI test items. The chosen LRSM 3 exhibited a significant correlation of 0.887 with the observed difficulty of the EI test items. Factors influencing the EI test's difficulty included the number of syllables, vocabulary score, and content word density. While the impact of the number of syllables concurred with the findings of several prior studies (Graham et al., 2010: Hendrickson et al., 2010: Ortega et al., 2002: Yan et al., 2016: Kim et al., 2016: Campfield, 2017: Yan, 2020: David & Norris, 2021), the expected influence of the number of ecels—a factor tied to the characteristics of the Korean language—did not turn out to be significant. However, it exhibited a substantial correlation with the number of syllables.

Similar to Isbell and Son's (2022) study, the present study found vocabulary score to be an influencing factor. However, contrary to their observation (Isbell & Son, 2022) that the number of function words had a significant impact, this study did not find either the number of function words or grammar score to be a significant factor. Instead, content word density emerged as an additional influencing factor. The lack of significance of grammar-related factors in our study was probably because the items were created with low grammatical complexity to maintain authenticity during the modification process.

In contrast to Isbell and Son (2022), our study did not find the numbers of content words, function words, or morphemes to be significant factors influencing item difficulty. Rather, we found content word density to be significant. In Graham et al.'s (2010) study, density was significant and the number of morphemes was insignificant, whereas Campfield (2017) observed that the opposite to be true. Our findings align with those of Graham et al. (2010). Further, Additionally, the results suggest that item difficulty decreases as content words density increases. However, it is generally expected that higher lexical density leads to higher information load, making text reading more difficult. Nonetheless, due to the interplay of content word density and vocabulary scores, which include the lexical frequency and difficulty, such outcomes may have arisen, necessitating further research on this matter. In summary, our study underscores the need to consider various linguistic factors that affect item difficulty when designing EI tests. The findings particularly emphasize the significance of considering the number of syllables, the presence of challenging

vocabulary, and the proportion of content words among morphemes in the design of effective EI tests.

Conclusions

In this study, we developed, tested, and validated a Korean EI test. Drawing on previous research on EI test development and validation, we formulated three research questions. The first question aimed to explore the correlation between the EI test and the TOPIK, assessing its effectiveness as a proficiency measurement instrument. The second question focused on the facets of the Korean EI test—learner, item, rater, and construct—to determine discernible patterns. The third question investigated the linguistic factors that influence EI test performance.

The Korean EI test developed in this study employed an analytical scoring approach using dual constructs—content and delivery—on a 0–5-point scale. Our exploration of the first research question involved a comprehensive analysis of correlations and ANOVA outcomes, which unveiled a significant positive correlation between TOPIK scores and the EI test's average scores. Moreover, when demarcating proficiency levels based on TOPIK scores, the Korean EI test exhibited a pronounced ability to differentiate between beginner and intermediate groups, as well as beginner and advanced groups. However, this ability was less pronounced when differentiating between intermediate and advanced groups. In addressing our second research question, a rigorous MFRM analysis demonstrated that most learners taking the EI test did not experience significant challenges and confirmed consistent and reliable scoring among all raters. These findings underscore the relatively straightforward nature of ensuring the EI test's reliability.

Our investigation into item difficulty and comparison with similar studies that have employed distinct item modifications highlighted the multifaceted factors influencing item difficulty. Beyond the number of syllables, grammatical intricacies and lexical characteristics had notable impacts on the difficulty of items. Notably, while the delivery construct exhibited stricter scoring than the content construct, it was difficult to discern the nuances of this difference due to the prevalence of high scores (4–5 points) across both constructs. Consequently, it is imperative to reevaluate the implications of analytically scored content and delivery constructs, drawing insights from learners who are faced with more difficult challenges in the developed test. Addressing our third research question, the EIRM analysis indicated that the number of syllables, vocabulary score, and content word density are pivotal factors when devising models to comprehensively determine the difficulty of an EI test.

This study's key contribution is its transformation of the existing items from Ortega et al. (2002) into authentic Korean expressions and its rigorous testing with a large number of learners and raters. The study also comprehensively addressed three research questions related to the validation of EI tests and was strengthened by the use of robust analysis methodologies. However, given the current trend of testing emphasizing consequential validity and washback, as well as communicative language testing, there is a potential negative washback effect of EI. Students might engage more in practicing listening and repeating for test preparation rather than actively participating in conversations (Cox & Davies, 2012). Therefore, caution is required when using EI in classroom

assessments. The utility of EI can be confirmed in the context of emphasizing communicative activities and interaction by using it alongside other communicative tasks (Yan et al., 2016). Furthermore, considering that individuals frequently incorporate their interlocutor's speech and grammar into their own during actual conversations, EI tasks can be deemed quite realistic (Van Moere, 2012).

Abbreviations

EI	Elicited Imitation
EIRM	Explanatory Item Response Modeling
IRT	Item Response Theory
L1	First language
L2	Second language
LRSM	Linear Rating Scale Model
MFRM	Many Facets Rasch Model
NIA	National Information Society Agency
NIKL	National Institute of the Korean Language
PCM	Partial Credit Model
RSM	Rating Scale Model
TOPIK	Test of Proficiency in Korean

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40468-024-00287-z.

Additional file 1: Appendix 1

Acknowledgements
This study was conducted using AI learning data "Korean voice data of Western and Asian language users for language education". These data were derived from the National Information Society Agency's project to build data for artificial intelligence learning, organized by Korea's Ministry of Science and ICT. This paper was presented at the 28th Annual Conference of the American Association of Teachers of Korean (AATK).

Authors' contributions
HK made substantial contributions to the study's conception, test design, and data interpretation. CS contributed to the test design, data acquisition and analysis, and data interpretation. JK was actively involved in data acquisition, data analysis, and data interpretation. HJ contributed to the test design, data acquisition, and data interpretation. JP contributed to the test design, data acquisition, and data analysis. All authors significantly contributed to the manuscript and have read and approved the final version.

Funding
The data used in this research were collected as part of a support project overseen by the National Information Society Agency in Korea that aims to collect artificial intelligence learning data. Direct research grants were not provided for this study. Parties external to the research had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.
This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2023S1A5A2A01082684). The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The datasets used and analyzed during the current study are available from the National Information Society Agency on reasonable request.

Declarations

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

Received: 23 October 2023 Accepted: 11 April 2024
Published online: 06 May 2024

References

Akbary, M., Benzaia, L. A., Jarvis, S., & Park, H. I. (2023). Evaluating the utility of elicited imitation as a measure of listening comprehension in the context of forensic linguistics. *Research Methods in Applied Linguistics, 2*(3), 100067. https://doi.org/10.1016/j.rmal.2023.100067

Baker, F. B. (2001). The Basics of Item Response Theory (2nd ed.). ERIC.
Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second-language competence. In E. E. Tarone, S. M. Gass, & A. D. Cohen (Eds.), *Research methods in second-language acquisition* (pp. 245–261). Lawrence Erlbaum Associates.
Bowden, H. W. (2016). Assessing second-language oral proficiency for research: The Spanish elicited imitation task. *Studies in Second Language Acquisition, 38*(4), 647–675. https://doi.org/10.1017/S0272263115000443
Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge: What can heritage language learners contribute? *Studies in Second Language Acquisition, 33*(2), 247–271. https://doi.org/10.1017/S0272263110000756
Bulut, O. (2021). eirm: Explanatory item response modeling for dichotomous and polytomous item responses (R package version 0.3.0) [Computer software]. https://doi.org/10.5281/zenodo.4556285, https://CRAN.R-project.org/package=eirm
Bulut, O., Gorgun, G., & Yildirim-Erbasli, S. N. (2021). Estimating explanatory extensions of dichotomous and polytomous Rasch models: The eirm package in R. *Psych, 3*(3), 308–321. https://doi.org/10.3390/psych3030023
Burger, S., & Chrétien, M. (2001). The development of oral production in content-based second language courses at the University of Ottawa. *Canadian Modern Language Review, 58*(1), 84–102. https://doi.org/10.3138/cmlr.58.1.84
Campfield, D. E. (2017). Lexical difficulty–using elicited imitation to study child L2. *Language Testing, 34*(2), 197–221. https://doi.org/10.1177/0265532215623580
Chaudron, C., Prior, M., & Kozok, U. (2005, July). Elicited imitation as an oral proficiency measure. In 14th World Congress of Applied Linguistics, Madison, WI.
Cox, T., & Davies, R. S. (2012). Using automatic speech recognition technology with elicited oral response testing. Calico Journal, 29(4), 601–618. https://www.jstor.org/stable/calicojournal.29.4.601
Davis, L., & Norris, J. (2021). Developing an innovative elicited imitation task for efficient English proficiency assessment. *ETS Research Report Series, 2021*(1), 1–30. https://doi.org/10.1002/ets2.12338
Deygers, B. (2020). Elicited imitation: A test for all learners? Examining the EI performance of learners with diverging educational backgrounds. *Studies in Second Language Acquisition, 42*(5), 933–957. https://doi.org/10.1017/S0272263120000082X
Eckes, T. (2015). *Introduction to many-facet Rasch measurement* [2nd Edition]. Peter Lang.
Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*(2), 141–172. https://doi.org/10.1017/S0272263105050096
Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27*(3), 464–491. https://doi.org/10.1093/applin/aml001
Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning, 66*(2), 419–447. https://doi.org/10.1111/lang.12157
Graham, C. R., Lonsdale, D., Kennington, C., Johnson, A., & McGhee, J. (2008). Elicited imitation as an oral proficiency measure with ASR scoring. Proceedings of the 6th International Conference on Language Resources and Evaluation, Paris, France. European Language Resources Association. https://www.researchgate.net/profile/Casey-Kennington-2/publication/220746202_Elicited_Imitation_as_an_Oral_Proficiency_Measure_with_ASR_Scoring/links/00b4953198ab59df55000000/Elicited-Imitation-as-an-Oral-Proficiency-Measure-with-ASR-Scoring.pdf
Graham, C. R., McGhee, J., Millard, B., Prior, M., Watanabe, Y., & Lee, S. (2010). The role of lexical choice in elicited imitation item difficulty. Selected proceedings of the 2008 second language research forum. Somerville, MA: Cascadilla Proceedings Project. https://www.lingref.com/cpp/slrf/2008/paper2385.pdf
Han, S. (2014). The status and future directions of Korean language proficiency assessment- a focus on evaluation of proficiency in speaking and writing with in the context of Korean language proficiency assessment. *The Fifth International Conference on Korean Language Education*, *2014*, 49–53.
Harsch, C. (2014). General language proficiency revisited: Current and future issues. *Language Assessment Quarterly, 11*(2), 152–169. https://doi.org/10.1080/15434303.2014.902059
Hendrickson, R., Aitken, M., McGhee, J., Johnson, A., Prior, M., Watanabe, Y., & Lee, S. (2010). What makes an item difficult? A syntactic, lexical, and morphological study of elicited imitation test items. Selected Proceedings of the 2008 Second Language Research Forum, Somerville, MA: Cascadilla Proceedings Project. bit.ly/3TWup4W
Heo, H., & Lee, Y. (2010). The relationship between the performance of sentence repetition and sentence production in school-age children. Phonetics and Speech Sciences, 2(1), 127–133. https://koreascience.kr/article/JAKO201019455946045.pdf
Isbell, D. R., & Son, Y. (2022). Measurement properties of a standardized Elicited Imitation Test: An integrative data analysis. *Studies in Second Language Acquisition, 44*(3), 859–885. https://doi.org/10.1017/S0272263121000383
Iwashita, N. (2006). Syntactic complexity measures and their relation to oral proficiency in Japanese as a foreign language. *Language Assessment Quarterly: An International Journal, 3*(2), 151–169. https://doi.org/10.1207/s15434311laq0302_4
Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review, 64*(1), 215–238. https://doi.org/10.3138/cmlr.64.1.215
Kim, Y., Tracy-Ventura, N., & Jung, Y. (2016). A measure of proficiency or short-term memory? Validation of an Elicited Imitation Test for SLA research. *The Modern Language Journal, 100*(3), 655–673. https://doi.org/10.1111/modl.12346
Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3–31. https://doi.org/10.1191/0265532202lt218oa
Kostromitina, M., & Plonsky, L. (2022). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition, 44*(3), 886–911. https://doi.org/10.1017/S0272263121000395
Lawson, D. M., & Brailovsky, C. (2006). The presence and impact of local item dependence on objective structured clinical examinations scores and the potential use of the polytomous, many-facet Rasch model. *Journal of Manipulative and Physiological Therapeutics, 29*(8), 651–657. https://doi.org/10.1016/j.jmpt.2006.08.002
Lee, J. (2005). A study on the sentence repetition performance between normal children and children with language-delayed. MD dissertation, Daegu Univertity

Lee, H. .S. (2008). Effects of using Mani-Faceted Rasch Model on accuracy of essay scoring under various scoring designs and rater characteristics. *Journal of Educational Evaluation, 21*(4), 129–152. G704–000051.2008.21.4.005.
Linacre, J. M. (2016). Winsteps® Rasch measurement computer program [Computer software]. https://www.winsteps.com/facets.htm
Lonsdale, D. W., & Chritensen, C. (2011). Automating the scoring of elicited imitation tests. In Symposium on Machine Learning in Speech and Language Processing. https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=7747&context=facpub
Lunz, M. E., & Stahl, J. A. (1990). Judge consistency and severity across grading periods. *Evaluation & the Health Professions, 13*(4), 425–444. https://doi.org/10.1177/016327879001300405
McDade, H. L., Simpson, M. A., & Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar: A question of validity. *Journal of Speech and Hearing Disorders, 47*(1), 19–24. https://doi.org/10.1044/jshd.4701.19
McManus, K., & Liu, Y. (2022). Using elicited imitation to measure global oral proficiency in SLA research: A close replication study. *Language Teaching, 55*(1), 116–135. https://doi.org/10.1017/S026144482000021X
Millard, B. J. (2011). Oral proficiency assessment of French using an elicited imitation test and automatic speech recognition. Brigham Young University.
National Institute of the Korean Language. (2003). A Korean vocabulary list for learners. Retrieved from https://www.korean.go.kr/front/etcData/etcDataView.do?mn_id=46&etc_seq=71
National Institute of the Korean Language. (2017). Applied Research for the International Standard Curriculum of Korean Language Retrieved from https://www.korean.go.kr/front/reportData/reportDataView.do?mn_id=207&report_seq=932
Norris, J. M., Sasayama, S., & Kim, M. (2023). The relationship between poststimulus pause, learner proficiency, and working memory in an Elicited Imitation Task. *Studies in Second Language Acquisition, 45*(5), 1370–1387. https://doi.org/10.1017/S0272263122000274
Oh, D. Y., Yim, D., & Yim, D. (2013). Non-word repetition and sentence repetition performance in 2–3 years old late talkers and normal children. Communication Sciences & Disorders, 18(3), 277–287. https://doi.org/10.12963/csd.13053
Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October). An investigation of elicited imitation tasks in crosslinguistic SLA research. Second language research forum, Toronto.
Park, H. I., Solon, M., Henderson, C., & Dehghan-Chaleshtori, M. (2020). The roles of working memory and oral language abilities in elicited imitation performance. *The Modern Language Journal, 104*(1), 133–151. https://doi.org/10.1111/modl.12618
Serafini, E. J. (2013). Cognitive and psychosocial factors in the long-term development of implicit and explicit second language knowledge in adult learners of Spanish at increasing proficiency. Georgetown University.
Solon, M., Park, H. I., Dehghan-Chaleshtori, M., Carver, C., & Long, A. Y. (2022). Exploring an elicited imitation task as a measure of heritage language proficiency. *Studies in Second Language Acquisition, 44*(4), 1095–1123. https://doi.org/10.1017/S0272263121000905
Solon, M., Park, H. I., Henderson, C., & Dehghan-Chaleshtori, M. (2019). Revisiting the Spanish elicited imitation task: A tool for assessing advanced language learners? *Studies in Second Language Acquisition, 41*(5), 1027–1053. https://doi.org/10.1017/S0272263119000342
Spada, N., Shiu, J. L. J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning, 65*(3), 723–751. https://doi.org/10.1111/lang.12129
Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning, 65*(4), 860–895. https://doi.org/10.1111/lang.12138
Thompson, C. A. (2013). The development and validation of a Spanish elicited imitation test of oral language proficiency for the Missionary Training Center. Brigham Young University.
Tracy-Ventura, N., McManus, K., Norris, J. M., Ortega, L., & Leclercq, P. (2014). Repeat as much as you can": Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmon, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143–166). Multilingual Matters.
Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325–344. https://doi.org/10.1177/0265532211424478
Vinther, T. (2002). Elicited imitation: A brief overview. *International Journal of Applied Linguistics, 12*(1), 54–73. https://doi.org/10.1111/1473-4192.00024
Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263–287. https://doi.org/10.1177/026553229801500205
West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition, 4*(3), 203–222. https://doi.org/10.1515/langcog-2012-0011
Wu, S., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals, 46*(4), 680–704. https://doi.org/10.1111/flan.12063
Wu, S., Tio, Y. P., & Ortega, L. (2022). Elicited imitation as a measure of L2 proficiency: New insights from a comparison of two L2 English parallel forms. *Studies in Second Language Acquisition, 44*(1), 271–300. https://doi.org/10.1017/S0272263121000103
Wu, S., Tio, Y. P., & Zhao, Y. (2023). Examining the comparability of parallel English and Chinese elicited imitation tasks. *Research Methods in Applied Linguistics, 2*(3), 100058. https://doi.org/10.1016/j.rmal.2023.100058
Yan, X. (2020). Unpacking the relationship between formulaic sequences and speech fluency on elicited imitation tasks: Proficiency level, sentence length, and fluency dimensions. *Tesol Quarterly, 54*(2), 460–487. https://doi.org/10.1002/tesq.556
Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing, 33*(4), 497–528. https://doi.org/10.1177/0265532215594643

Publisher' s Note