Using LLMs for Item Creation: Validating the Potential of Automatically Generated Sentence Repetition Test Items for Language Assessment

Sarah Löber[1,2], Björn Rudzewitz[1,2], Yuan Chu[3], Mengyuan He[3], Shiqin Liu[3], Yushan Ye[3], Xiaobin Chen[1,2]
[1]Hector Research Institute of Education Sciences and Psychology,
University of Tübingen, Germany
[2]LEAD Graduate School and Research Network, University of Tübingen, Germany
[3]Guangzhou City University of Technology, China
{sarah.loeber,bjoern.rudzewitz,xiaobin.chen}@uni-tuebingen.de
{chuyuan,hemy,liusq,yeys}@gcu.edu.cn

Abstract

Various aspects of the Elicited Imitation Test (EIT), a sentence repetition task for language assessment, can be automated, for example in terms of test administration or automatic scoring. It is potentially also possible to generate test items with Large Language Models (LLMs). This study investigates the potential of GPT-4o for item creation in the context of EIT, creating a parallel form to two popular and validated tests. We analysed the tests in terms of their linguistic and psychometric properties. While the items created by the LLM show some difference in grammatical structures when compared to human-written items, linguistic complexity results did not differ significantly between tests. Psychometric properties showed only minor differences. These findings lend support to the potential of Automatic Item Generation with LLMs in the context of sentence repetition tasks and might support the process of standardisation in SLA research and testing by enabling parallel test creation.

1 Introduction

Creating items for language tests places a particular kind of strain on the test developer: it requires resources such as personnel and time as well as expertise, making it a complex and labourious process. In the realm of language assessment, recent advancements have made possible the automation of this task. For example, Settles et al. (2020) used machine learning techniques for the automatic creation of proficiency assessments for adaptive testing. Likewise, Attali et al. (2022) utilised a transformers-based approach for generating reading comprehension assessments. As these examples make apparent, generating items automatically can accelerate the process of test creation and reduce the amount of labour typically associated with this task. Indeed, Automatic Item Generation (AIG) can be found in many contexts nowadays, from generating reading comprehension questions (Huang and He, 2016; Sayin and Gierl, 2024) to multiple choice items (Gierl et al., 2012). Another test format that lends itself well to automation and therefore AIG is the Elicited Imitation Test (EIT). While AIG is a promising and efficient approach, questions remain about construct representation and validity.

Traditionally, item creation for the EIT has been manual, relying on test creators for writing their own items and tests. Since EIT is a sentence repetition task, the sentences need to be carefully controlled in terms of length and linguistic complexity. Automating this process would reduce the amount of work related to item creation and increase the possibility of adaptive testing in the context of EIT. Furthermore, AIG has the potential to tackle another problem: the lack of standardisation in EITs and in language research generally (Isbell and Son, 2022), which becomes problematic when researchers want to compare a construct, e.g. language proficiency, across studies or when the need for repeated assessment arises. Parallel and comparable forms are a solution to this issue, provided they demonstrate comparable psychometric properties.

As suggested in recent literature, pre-trained LLMs and prompting are increasingly used for engineering test items (Song et al., 2025). One clear advantage of this approach is that using pre-trained LLMs is an accessible option for most people without a strong computational background that is otherwise needed to engineer systems for automatic item generation. However, questions remain about the quality and linguistic properties of these generated items, particularly in the context of validity in

Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications, pages 443–454

(language) testing.

The aims of this study are to (1) investigate the linguistic properties of automatically generated items with LLMs in the context of a parallel form of an EIT, and (2) to examine if these test items display validity evidence in order to support the development of standardised and comparable EIT forms.

2 Theoretical background

2.1 Elicited Imitation

The Elicited Imitation Test (EIT) is a sentence repetition task, commonly used as a measure of general language proficiency. The test is also used in Second Language Acquisition (SLA) research, for example, to measure implicit grammatical knowledge of a second language (Erlam, 2006). Furthermore, the EIT can also be used as a placement test (Yan et al., 2020) or as a diagnostic tool for language acquisition in children (Klem et al., 2015).

The procedure of an EIT is as follows: The test taker hears sentences one by one, which they have to repeat verbatim. There is usually a short pause between the stimulus and the repetition to avoid over-reliance on working memory. The rationale behind sentence repetition is that the test taker reconstructs the meaning of the sentence based on their own grammatical knowledge.

Various design implementations can be employed when designing an EIT, for example using sentences of varying length (Ortega et al., 2002), including a distractor instead of a pause (Erlam, 2006), including ungrammatical sentences, or including only items that target a specific construct such as the passive-form (Spada et al., 2015).

So far, most of the test creation in the context of the EIT has been manual. Items have been written by researchers or test developers, which is a labour-intensive and complex task. Furthermore, it makes scaling of the test for parallel testing or testing on a larger scale difficult. This can cause problems as comparable testing and standardisation are two things that are currently lacking in SLA research (Isbell and Son, 2022). This difficulty could be tackled by automating EI test generation and administration to increase its scalability.

Research efforts have concentrated on automating the EIT using various technologies (McGuire and Larson-Hall, 2025; Isbell et al., 2023; Kim et al., 2024). For instance, Isbell et al. (2023) used Automatic Speech Recognition (ASR) and string matching to automate the scoring of the Korean EIT, achieving very high correlations (> .90) between human and automated scores. Automatic scoring has also been implemented for written EITs (Chifligarov et al., 2025). Furthermore, the EIT can be digitised and taken on the web, yielding results highly similar to lab-based approaches (Kim et al., 2024). In addition to these two areas, researchers have also investigated item creation for EITs (Christensen et al., 2010). Unlike delivery and scoring, EIT item creation has largely been manual.

2.2 Item Creation Approaches in EIT

Items used in existing EITs have traditionally been hand-crafted. For example, the most well-known EIT by Ortega et al. (2002) used hand-crafted items, as do the parallel forms in several languages. In order to create parallel English versions for comparability, Wu et al. (2022) also hand-crafted items based on the original Ortega EIT. Using appropriate items in the context of EIT is important, since the purpose of the test can vary from testing advanced learners (Solon et al., 2019) by including more complex items to testing specific grammatical target structures, e.g., *passives*, (Godfroid and Kim, 2021; Spada et al., 2015) to see if learners have acquired those structures.

To our knowledge, the first study using a corpus-based item bank in the context of EIT was carried out by Christensen et al. (2010). In this study, the authors used an annotated corpus for flexible item selection and then compared an existing EIT, using human-written items, with an EIT created with their automatic item selection tool. Interestingly, the automatically created version of their EIT had even higher correlation scores with an external measurement used in their study, the Spoken Language Achievement Test (SLAT), than the manually created EIT. While this is promising for the generation of items, the finding might also hint at different behaviour or validity characteristics of the two forms.

Using a corpus in the context of creating EIT items holds advantages: sentences from a corpus can be more authentic than items written for EIT use and, additionally, choosing sentences from a corpus is also less work than hand-crafting EIT items. However, this approach can also introduce a substantial amount of noise and effort: sentences can be out of context or inappropriate. Furthermore, items might be too long or overly complicated and require manual adaptation. A corpus might also be limited to a specific domain only,

which might be unfamiliar to some population of learners. Another difficulty is having control over potential target structures in the corpus sentences, which requires additional annotation of the corpus. A corpus-based item bank in the context of EIT therefore also places demands on test creators and does not entirely simplify the item creation process.

In summary, previous approaches in EIT relied on manual creation or corpus-based retrieval of items. Corpus-based methods are a step towards automation, however, they only offer a partial solution to the problems of scalability, effort, structural control and standardisation. These limitations encourage the exploration of alternative approaches, such as Automatic Item Generation (AIG).

2.3 Automatic Item Generation

AIG refers to the creation of (test) items using automated approaches such as machine learning and carries several advantages, among them more time- and resource-efficient item creation (Circi et al., 2023).

This approach finds application in several fields and contexts, for example, the creation of multiple choice items in a medical context (Gierl et al., 2012) or for generating reading comprehension questions (Huang and He, 2016; Sayin and Gierl, 2024). For generating items, cognitive models are employed, which model the domain knowledge, content, or skills needed to develop items (Gierl et al., 2012). Recent advancements in Large Language Models (LLMs) have created new approaches in AIG. While AIG usually involves the complex task of incorporating cognitive models in item creation, LLMs have enabled a less technically complex approach, involving few- or zero-shot prompting (Attali et al., 2022) and allowing for more flexibility and fewer constraints in the item creation process, for example by allowing the creation of a variety of item types in various subject domains (Tan et al., 2024). Generally, using prompt engineering with pre-trained models is a growing trend in the field (Song et al., 2025) and is additionally a feasible method for automatic item creation for people without a computational background, for example teachers and educators.

Beyond the generation of items, an equally important aspect in AIG is the evaluation and validation of test items. In the context of AIG with LLMs, item and measurement properties are often not reported or evaluated (Tan et al., 2024). Especially in the context of high-stakes tests, validation and testing of items is crucial to make sure that the test actually tests what it is supposed to test. This is also important in the case of the EIT, as generated items may differ in linguistic structures or lengths when compared to established test items, potentially altering cognitive load for the participant and item difficulty.

Tests are often evaluated in terms of validity. Broadly speaking, validity can be defined as the "meaning of test scores" (Messick, 1995, p. 5), referring to the overall concept as describing whether a test measures what it intends to measure. Validity can be measured in several ways. An example is construct validity, an aspect of validity that tests if a test indeed measures the underlying construct it is trying to measure, or criterion validity, that measures how related test results are to another test (El-Hamamsy et al., 2022).

For the evaluation of AIG items, recent work has focused on automated approaches, but using a human-in-the-loop evaluation of items, for example, on the rating of the quality of generated items (Kim et al., 2025; Ma et al., 2025). Another possibility for assessing the suitability of items are psychometric methods such as Item Response Theory (IRT), which can inform about validity and quality in the context of AIG (Falcão et al., 2023).

Using AIG in the context of Elicited Imitation Tests certainly holds potential: it enables parallel testing, the generation of a large item bank, and supports the labour-intensive process of crafting items. While an earlier approach (Christensen et al., 2010) used corpora and annotation to find suitable items, LLMs can simplify this process by responding flexibly to certain demands, e.g. the presence of specific constructs or topics. This allows for more control than a corpus-based setup.

3 Present Study

Taken together, AIG seems to be promising in the context of EITs, allowing for flexibility and opening potential for the creation of parallel forms, which in turn can help with the process of standardisation. However, automatically generated items would need to be validated in order to make claims about their potential to be used in language assessment. To date, no study has systematically investigated whether LLM-generated items can be used as valid EIT items. To research this, we use an LLM-generated EIT, employing GPT-4o. We opted for GPT as it is the most widely adopted tool in ed-

ucation (Bhullar et al., 2024), accessible through a web interface and frequently used in AIG (Tan et al., 2024; Chan et al., 2025), which helps with comparability of the findings. This way, our findings are directly relevant to educators and students in practice. For comparability, we will concentrate on parallel forms in particular. Since grammatical structures and complexity are of specific interest to EIT items, we will investigate item properties and measurement properties in two ways: evaluating items and tests in terms of their linguistic and psychometric properties. The research questions are as follows:

1. To what extent do the properties of GPT-generated test items differ from those of established test items? More specifically,
   (a) Are they different in terms of linguistic complexity?
   (b) Are they different in terms of grammatical structures used?
2. How do the automatically generated tests compare to established tests in terms of validity?

4 Method

4.1 Materials

Three EITs were used in this study: (1) An established EIT by Ortega et al. (2002), (2) the parallel version by Wu et al. (2022) and (3) an EIT generated by GPT-4o. The items of the GPT-generated EIT can be found in Appendix A.

Our automatic item generation approach was as follows: we prompted GPT-4o to create an EIT with 30 items, varying in length, starting at 7 syllables. We employed few-shot prompting for automatic item generation, including the items of the original EIT by Ortega et al. (2002) as well as their syllable count as examples. The model output, 30 items, was used for the resulting GPT-generated EIT. The prompt can be found in Appendix C.

The tests were all uploaded to ILAP (Löber et al., 2024), an assessment platform, where participants could take the tests on their own devices via a web browser. All EIT tests used the same procedure: The participants saw instructions about the test, followed by three practice items that were not counted towards the final score. Then, a notification about the start of the test followed. The test then began. After each sentence, there was a beep followed by a four-second pause. Participants then had eight seconds to repeat the sentence they had just heard.

Audio stimuli for all items on the three EITs were created with the Amazon AWS text-to-speech service, which had been integrated into the platform.

Furthermore, we administered questionnaires about participants' demographic background and a C-Test as a measure for general English language proficiency taken from Ishihara et al. (2003). The C-Test was programmed with jsPsych (De Leeuw, 2015) in JavaScript and integrated into the platform used for EIT data collection. The test consisted of three text passages with half of every fifth or sixth word removed. The maximum score on the C-test was 50 points (15 points for each of the first 2 passages and 20 points for the last passage).

The EITs were scored automatically, using the ASR and string matching approach by Isbell et al. (2023). We used accuracy based on Mean Error Rate (MER) - (100 - MER) - as the scoring metric (McGuire and Larson-Hall, 2025). We then mapped our continuous scores to ordinal scores for better comparability with the ordinal scoring scheme by Ortega et al. (2002), encompassing scores from 0-4. We followed the approach taken by Isbell et al. (2023) for mapping the scores, which can be found below in Table 1. For total scores, we took the sum of all item scores. The maximum score on the EITs was 120.

Ordinal score	Continuous score
4	91 - 100
3	71 - 90
2	41 - 70
1	21 - 40
0	20 or below

Table 1: The ordinal scoring scheme, mapping continuous scores to ordinal scores

The C-Test was also automatically scored, assigning binary scores (1 or 0) to each gap. This was done by employing string matching between the participants' response to the expected response for each gap. We then used the sum of all answers as the total score. The maximum score on the C-Test was 50.

4.2 Study Design

The study ran for four weeks. Each week, a teacher administered an EIT to a class in a randomised order, e.g. class 1 completing EIT 1 in week 1, EIT 2 in week 2 and EIT 3 in week 3 while class 2

completed EIT 2 in week 1, EIT 3 in week 2 and EIT 1 in week 3, etc.

Students created an individual accounts on the platform used for testing. The EITs were implemented into the platform by the researchers, where they could be administered to the students with an access code. Before being able to access any test, students were shown a consent form where they could indicate whether they consented or not to their data being used for study purposes. Furthermore, we also implemented the questionnaire and C-Test on the platform. These could be taken by students at any time during the course of the 4-week study.

4.3 Participants

Participants were 73 undergraduate students from a university in China, majoring in English. 60 of the students were female and the mean age was 20.1 years (SD = 1.55), ranging from 18 to 25 years of age. All students reported Chinese as their native language and the majority (72%) reported English as their second language. Students' self-reported English proficiency was intermediate to upper intermediate.

4.4 Analysis

Data were analysed with R (version 4.5.0), using the packages stringr (Wickham, 2019), lme4 (Bates et al., 2015), lavaan (Rosseel, 2012), mirt (Chalmers, 2012) and koRpus (Michalke et al., 2021). We also analysed all test items in terms of the grammatical structures and corresponding learner levels with the rule-based annotation tool POLKE (Sagirov and Chen, 2025). POLKE annotates text with grammatical structures and their corresponding Common European Framework of References (CEFR) (Council of Europe, 2001) learner levels (A1-C2) and is based on the English Grammar Profile (O'Keeffe and Mark, 2017). In addition to the grammatical structures and learner levels, the tool also supplies super-categories of structures (e.g. nouns, verbs, adjectives) as well as sub-categories (e.g. types, linking, phrasal).

In addition, we analysed the EIT items in terms of their linguistic complexity, using the tool CTAP (Chen and Meurers, 2016) as well as the stringr and koRpus packages in R. For measure selection, we followed Bulté et al. (2025)'s summary of core and noncore measures of complexity, selecting six measures. We also analysed word frequency, using the SUBTLEXus log-10 frequency of all tokens. The full overview of complexity measures used can be found in Appendix B.

Furthermore, we employed Item Response Theory (IRT) and Confirmatory Factor Analysis (CFA) as well as correlations to analyse the test properties of the 3 EITs.

5 Results

5.1 Descriptive Results

Not all participants completed every measure. Specifically, 68 participants finished the background questionnaire, 62 the C-Test, 63 EIT 1, 61 EIT 2 and 62 EIT 3, with 53 participants completing all 3 EITs. We analysed all data available to us, which meant that for example, the correlation analyses were conducted with only participants who completed all measures included in the analysis.

The mean score on the C-Test was 38.96 (SD = 11.01). Mean scores on the EITs were similar: EIT 1 (59.6, SD = 34.50) and EIT 3 (57.82, SD = 31.20) displayed highly similar scores, whereas EIT 2 (51.58, SD = 35.78) showed a slightly lower mean score. EIT scores showed a high range, 11-113 for EIT 1, 5-112 for EIT 2 and 5-113 for EIT 3. The tests themselves displayed high levels of similarity: The mean syllable length was between 10.4 and 10.97 syllables for all three tests and the mean sentence length was between 14.93 and 15.2 for all three tests. Descriptive statistics for the scores on the three tests, including standard deviations (SD), can be found in Table 2 and 3 below.

Test	n	Mean score	SD
Ortega (2002)	63	59.58	34.50
Wu et al. (2022)	61	51.58	35.78
GPT-4o	62	57.82	31.20

Table 2: Number of participants, mean score and standard deviation (SD) for the three tests.

Test	Mean SL	SD	Mean SC	SD
Ortega (2002)	10.83	2.59	14.93	3.34
Wu et al. (2022)	10.97	2.62	15.03	3.78
GPT-4o	10.40	2.62	15.20	3.89

Table 3: Descriptive statistics for test properties. SL = sentence length, SC = syllable count.

5.2 Items

We first analysed the POLKE annotations in terms of the distribution of super-categories. We excluded

categories with counts < 5 in order to meet the assumptions of chi-square tests. These were the categories future, passives, questions and reported speech. The distribution of the remaining categories differed significantly between tests, $\chi^2(22) = 39.839$, $p = 0.011$. Following this, we inspected the residuals of the chi-square test. The GPT-generated tests produced fewer structures belonging to the adverbs and present category and more structures belonging to the preposition category than EIT 1 and EIT 2. Figure 1 shows the distribution of super-categories.

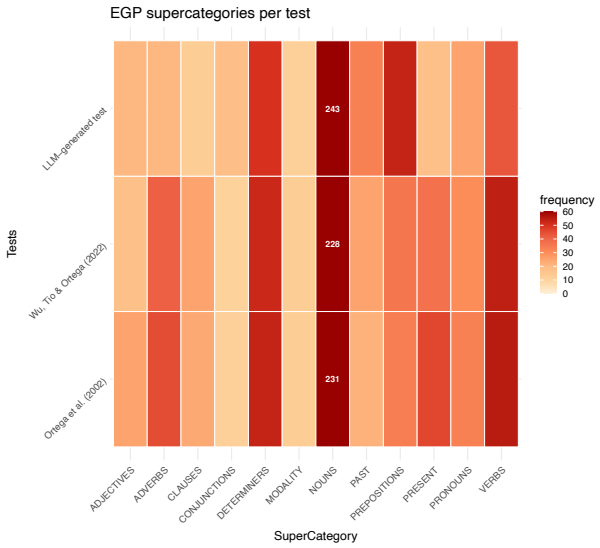

Figure 1: The distribution of super-categories for all three tests.

After that, we analysed whether there were differences in the distribution of CEFR learner levels within the POLKE-annotated grammar structures between the tests. The distribution of learner levels did not significantly differ between the tests, $\chi^2(10) = 9.8922$, $p = 0.45$. Figure 2 shows this distribution.

We then divided the complexity measures into sub-measures of syntactic (Table 4) and lexical complexity (Table 5). The results for each can be found in the tables below.

Prior to statistical testing, we used Shapiro-Wilk Tests to check for the distribution of data. If the tests were significant, we rejected the assumption of normality and used non-parametric tests. We first ran Kruskal-Wallis tests on the syntactic complexity measures. Three measures did not differ significantly between conditions: mean words per phrase ($p = 0.09$), mean phrases per clause ($p = 0.70$) and mean clauses per T-Unit ($p = 0.09$). One

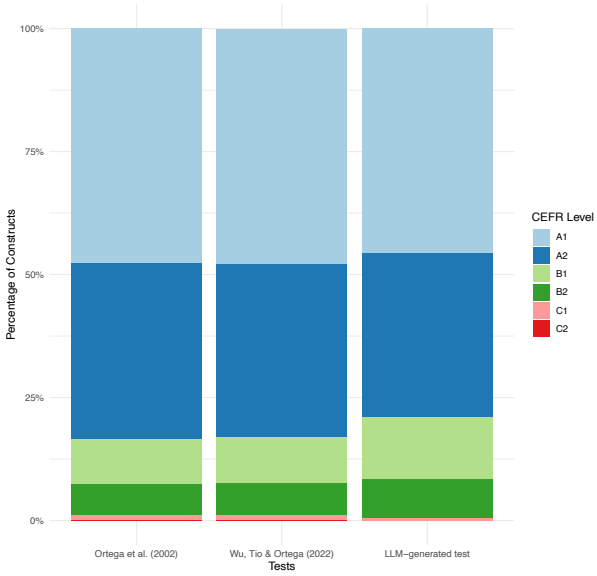

Figure 2: The distribution of learner levels associated with grammatical structures for all three tests, in percentages.

difference was significant: the normalised rate of occurrence of dependent clauses differed significantly between the tests, H(2) = 6.09, $p = 0.048$. We used Dunn's post hoc test with Bonferroni corrections to assess which groups differed. However, after adjusting, there was no significant difference between the Ortega et al. (2002) test and the GPT-generated form ($p = 0.177$) or between the Wu et al. (2022) form and the GPT-generated form ($p = 0.062$).

For lexical complexity, a Kruskal-Wallis test showed that the mean word length did not vary significantly between tests, H(2) = 2.93, $p = 0.231$. Word frequency also did not differ significantly between tests ($p = 0.178$).

Test	**MW phrase**	**MP clause**
Ortega (2002)	1.47	4.17
Wu et al. (2022)	1.59	3.98
GPT-4o	1.68	4.20
Test	**MC T-Unit**	**N. ROC**
Ortega (2002)	1.73	0.44
Wu et al. (2022)	1.76	0.47
GPT-4o	1.42	0.27

Table 4: Syntactic complexity mean scores for all three tests. MW = mean words, MP = mean phrases, MC = mean clauses, N. ROC = Normalised rate of occurrence of dependent clauses.

Test	Mean WL	MATTR	WF
Ortega (2002)	3.95	0.92	4.71
Wu et al. (2022)	4.01	0.91	4.60
GPT-4o	4.35	0.92	4.55

Table 5: Lexical complexity mean scores for the three tests. MATTR = moving-average type-token ratio, WL = word length, WF = word frequency.

Test	Factor loadings
Ortega (2002)	0.89
Wu et al. (2022)	0.85
GPT-4o	0.93

Table 6: Standardised factor loadings per test.

5.3 Validity

All three tests displayed very high internal consistency, $\alpha = 0.98$.

We first calculated Spearman correlations of the scores between tests. EIT 1 and 3 showed a correlation coefficient of $\rho = 0.80$ (shared variance: 0.64), EIT 1 and 2 showed a correlation coefficient of $\rho = 0.76$ (shared variance: 0.58), and EIT 2 and 3 correlated with a correlation coefficient of $\rho = 0.76$ (shared variance: 0.58).

Using IRT, we then fitted a 1PL graded response model to the 3 EITs and used common-person linking to obtain discrimination values for the items. All items showed acceptable discrimination values higher than 0.39 (Popham, 2000). We compared the discrimination values between tests, which differed significantly, $H(2) = 6.61$, $p = 0.037$. Dunn's post hoc test with Bonferroni corrections showed that the discrimination values differed only between the Wu et al. (2022) and GPT-generated form, $p = 0.038$, with the GPT-generated form showing higher discrimination values, but not between the Ortega et al. (2002) and GPT-generated form ($p = 0.22$) or the Ortega et al. (2002) and Wu et al. (2022) form ($p = 1$).

Focusing on the test level next, we employed confirmatory factor analysis (CFA) to obtain standardised factor loadings for the tests. We created a one-factor CFA model, using Full Information Maximum Likelihood (FIML) for missing data, with all three tests loading on the same latent variable, namely proficiency. The model was just-identified ($df = 0$), so we could not assess model fit. We then compared the standardised factor loadings of the tests. All tests displayed strong and significant factor loadings, indicating that they are connected in a highly similar way to the construct proficiency. The factor loadings can be found in table 6.

To obtain a measure of criterion validity, we correlated participants' final scores on the EIT with their final scores on the C-Test using Spearman correlation. All EITs showed a weak to moderate correlation with the C-Test, $\rho = 0.26$ for EIT 1, $\rho = 0.35$ for EIT 2, and $\rho = 0.29$ for EIT 3. Using Fisher's r-to-z transformation, we tested if the correlations differed between tests. There was no significant difference between EIT 1 and EIT 2 ($p = 0.55$), EIT 2 and 3 ($p = 0.72$) or EIT 1 and 3 ($p = 0.81$).

5.4 Scores

To find out if participants' final test scores differed significantly between tests, we ran a mixed linear model, modeling participants as random effects to account for individual differences between users. The resulting linear mixed model with users as random effects showed a significant main effect of test between EIT 1 and EIT 2 scores ($p = 0.021$), and no difference between EIT 1 and EIT 3 ($p = 0.66$). Participants therefore had a significantly lower score on EIT 2 than on EIT 1.

6 Discussion and Conclusion

In this study, we explored the potential for automatic item generation for parallel forms with GPT-generated items in the context of Elicited Imitation. Our results are promising: the GPT-generated test behaved very similarly to the original Ortega et al. (2002) form as well as the parallel form (Wu et al., 2022), displaying highly similar linguistic properties, high internal reliability as well as similar factor loadings and construct validity.

There were slight differences between the tests (human-written vs GPT-generated) when it came to the grammatical constructs used: The GPT-generated test displayed a lower number of adverbs and present (tense) and higher numbers of prepositions than the other two tests. However, there was no difference between the three tests in the learner levels associated with the grammatical constructs, potentially an indicator for the non-significant difference in the difficulty of items between the tests. In other words, the differences found in the use of grammatical structures did not seem to contribute to the difficulty of the tests. We should also add that we did not specifically prompt for following

the grammatical constructs used closely. Interestingly, the different grammatical structures do not seem to influence item or test difficulty.

The complexity analysis showed that all three tests score very similarly on all measures. The only significantly different measure, normalised rate of occurrence of dependent clauses, turned out to be non-significant in the post hoc tests.

In summary, the GPT-generated test deviates from the original EIT and the human-written parallel form on a few grammatical structures and not on complexity measures, but the deviations on grammatical structures do not seem to add to overall test difficulty of the GPT-generated form or interfere with form equivalence. This finding should be further investigated in future studies, focusing on item and test difficulty on EITs. While some studies have already investigated this topic (Hendrickson et al., 2010; Campfield, 2017) and found effects depending on the sentence length or lexical complexity, future studies might want to explore that topic further.

A slightly different picture emerged when looking at our psychometric analysis. While the Ortega et al. (2002) and GPT-generated form seem highly similar, the Wu et al. (2022) form displayed some deviations from both forms: the IRT analysis showed lower item discrimination scores than for the GPT-generated form. In terms of scores, the parallel form from Wu et al. (2022) showed a lower mean score. This is in line with the findings from the parallel forms study (Wu et al., 2022), where the authors found an 8% score difference between the parallel form and the Ortega et al. (2002) form. Between those two tests, we found a score difference of 14%. Interestingly, the TTS voice used in our tests was the same across tests, pointing to no added difficulty introduced by speaker variability as Wu et al. (2022) suspected as one of the culprits for enhanced test difficulty. The lower score on this test must therefore be for reasons beyond speaker variability and the complexity measures employed in this study. We also reported less shared variance between the tests than in Wu et al. (2022). This finding might be due to the different population used in this study.

An unexpected finding also emerged from our data: the correlation of the C-Test with performance on the EIT was quite low (weak correlation of ρ = 0.26 and ρ = 0.29 on EIT 1 and EIT 3, respectively, and moderate (ρ = 0.35) on EIT 2), which is contrary to what other studies found (r = 0.69 in Davis and Norris, 2021, r = 0.50 in Spada et al., 2015). A C-Test counts as a measure of holistic proficiency, but due to the nature of the C-Test, which is more geared towards writing and reading, it can be argued that the sub-skills of listening and speaking are not included, which is what the EIT taps into. These two sub-skills might have been developed less strongly when compared to writing and reading for our participants, which might explain the weak correlation. In our case, the mean scores on the C-Test and the EITs also point to C-Tests being easier for our participants than the EIT.

Taken together, our findings show that an LLM-generated test holds potential in creating valid EIT items in a parallel forms context. This finding could enable the scaling of tests, making it possible to create more parallel forms with LLM-generated items. Furthermore, this can potentially help to tackle the problem of standardisation in language research (Isbell and Son, 2022), by enabling the creation – at least in the case of EIT – of comparable forms to existing EITs. For future studies, it would be interesting to see if our findings would be applicable across language contexts. While we only concentrated on English in this study, performance of LLMs might be less consistent beyond that and it is unclear if the performance of GPT can be carried over to other languages. Since all the available parallel and comparable EIT forms across languages are created manually (e.g. Wu and Ortega, 2013; Bowden, 2016), studies could shed more light on the performance of automatically generated items across languages.

7 Limitations

A few limitations of this research should be mentioned. First of all, the number of participants (n = 73) was quite low, leading to less generalisable results. Due to this low number of participants, we were not able to carry out more sophisticated analyses, e.g. Differential Item Functioning and our IRT analysis should be taken as preliminary. Having more participants at hand, and possibly a more heterogeneous group of participants, would undoubtedly shed more light on validity and parallel form equivalence.

Furthermore, due to the nature of our online experiment, we experienced a number of dropout cases and missing data from the background questionnaire.

As already pointed out in the previous section, we also did not compare performance across different LLMs. GPT is one of the most readily available options and widely used in the context of education (Doughty et al., 2024; Vanzo et al., 2025) and is also relevant to educators, teachers and test creators. However, the performance of other models would also be of interest for this strand of research.

We also only tested 30 LLM-generated sentences. While our results seem promising for parallel test creation, future research should test a larger number of items to evaluate if our findings are generalisable to different contexts, e.g. repeated or adaptive testing.

Lastly, we did not take the ability of models to produce certain grammatical constructs on demand into account. While we found that the GPT-generated test mostly showed strong similarities to the two human-written forms in terms of supercategories, we didn't explicitly prompt for it. Further research could concentrate on this ability of LLMs, as this is central to some EITs targeting specific grammatical constructs and can shed light on large language models' abilities to produce grammatical constructs.

Acknowledgements

This work was supported by the German Ministry of Education and Science (BMBF) under Grant number 01IS22076. We would like to thank the anonymous reviewers for their helpful comments.

References

Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and M Ben Bolker. 2015. Package 'lme4'. *convergence*, 12(1):2.

Pritpal Singh Bhullar, Mahesh Joshi, and Ritesh Chugh. 2024. Chatgpt in higher education-a synthesis of the literature and a future research agenda. *Education and Information Technologies*, 29(16):21501–21522.

Harriet Wood Bowden. 2016. Assessing second-language oral proficiency for research: The Spanish elicited imitation task. *Studies in Second Language Acquisition*, 38(4):647–675.

Bram Bulté, Alex Housen, and Gabriele Pallotti. 2025. Complexity and difficulty in second language acquisition: A theoretical and methodological overview. *Language Learning*, 75(2):533–574.

Dorota E. Campfield. 2017. Lexical difficulty – using elicited imitation to study child L2. *Language Testing*, 34(2):197–221.

R Philip Chalmers. 2012. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48:1–29.

Kuang Wen Chan, Farhan Ali, Joonhyeong Park, Kah Shen Brandon Sham, Erdalyn Yeh Thong Tan, Francis Woon Chien Chong, Kun Qian, and Guan Kheng Sze. 2025. Automatic item generation in various stem subjects using large language model prompting. *Computers and Education: Artificial Intelligence*, 8:100344.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity at COLING*. The International Committee on Computational Linguistics.

Mihail Chifligarov, Jammila Laâguidi, Max Schellenberg, Alexander Dill, Anna Timukova, Anastasia Drackert, and Ronja Laarmann-Quante. 2025. Automated scoring of a German written elicited imitation test. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 237–247, Vienna, Austria. Association for Computational Linguistics.

Carl Christensen, Ross Hendrickson, and Deryle Lonsdale. 2010. Principled construction of elicited imitation tests. In *Proceedings of the 7th conference on international language resources and evaluation (LREC)*.

Ruhan Circi, Juanita Hicks, and Emmanuel Sikali. 2023. Automatic item generation: foundations and machine learning-based approaches for assessments. In *Frontiers in Education*, volume 8, page 858273.

Council of Europe. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Larry Davis and John Norris. 2021. Developing an innovative elicited imitation task for efficient english proficiency assessment. *ETS Research Report Series*, 2021(1):1–30.

Joshua R De Leeuw. 2015. jspsych: A Javascript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47(1):1–12.

Jacob Doughty, Zipiao Wan, Anishka Bompelli, Jubahed Qayum, Taozhi Wang, Juran Zhang, Yujia Zheng, Aidan Doyle, Pragnya Sridhar, Arav Agarwal, and 1 others. 2024. A comparative study of AI-generated (GPT-4) and human-crafted MCQs in programming

education. In *Proceedings of the 26th Australasian Computing Education Conference*, pages 114–123.

Laila El-Hamamsy, María Zapata-Cáceres, Estefanía Martín Barroso, Francesco Mondada, Jessica Dehler Zufferey, and Barbara Bruno. 2022. The competent computational thinking test: Development and validation of an unplugged computational thinking test for upper primary school. *Journal of Educational Computing Research*, 60(7):1818–1866.

Rosemary Erlam. 2006. Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, 27(3):464–491.

Filipe Falcão, Daniela Marques Pereira, Nuno Gonçalves, Andre De Champlain, Patrício Costa, and José Miguel Pêgo. 2023. A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. *Advances in Health Sciences Education*, 28(5):1441–1465.

Mark J Gierl, Hollis Lai, and Simon R Turner. 2012. Using automatic item generation to create multiple-choice test items. *Medical Education*, 46(8):757–765.

Aline Godfroid and Kathy Minhye Kim. 2021. The contribution of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition*, 43(3):606—-634.

Ross Hendrickson, Meghan Aitken, Jeremiah McGhee, and Aaron Johnson. 2010. What makes an item difficult? A syntactic, lexical, and morphological study of elicited imitation test items. In *Selected Proceedings of the 2008 Second Language Research Forum*, pages 48–56. Cascadilla Proceedings Project Somerville, MA.

Yan Huang and Lianzhen He. 2016. Automatic generation of short answer questions for reading comprehension assessment. *Natural Language Engineering*, 22(3):457–489.

Daniel R. Isbell, Kathy Minhye Kim, and Xiaobin Chen. 2023. Exploring the potential of automated speech recognition for scoring the Korean Elicited Imitation Test. *Research Methods in Applied Linguistics*, 2(3):100076.

Daniel R. Isbell and Young-A Son. 2022. Measurement properties of a standardized elicited imitation test: An integrative data analysis. *Studies in Second Language Acquisition*, 44(3):859–885.

Kenji Ishihara, Elizabeth Hiser, and Tae Okada. 2003. Modifying C-test for practical purposes. *Doshisha Studies in Language and Culture*, 5(4):539–568.

Euigyum Kim, Seewoo Li, Salah Khalil, and Hyo Jeong Shin. 2025. STAIR-AIG: Optimizing the automated item generation process through human-AI collaboration for critical thinking assessment. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 920–930, Vienna, Austria. Association for Computational Linguistics.

Kathy Minhye Kim, Xiaoyi Liu, Daniel R Isbell, and Xiaobin Chen. 2024. A comparison of lab-and web-based elicited imitation: Insights from explicit-implicit L2 grammar knowledge and L2 proficiency. *Studies in Second Language Acquisition*, pages 1–22.

Marianne Klem, Monica Melby-Lervåg, Bente Hagtvet, Solveig-Alma Halaas Lyster, Jan-Eric Gustafsson, and Charles Hulme. 2015. Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental Science*, 18(1):146–154.

Sarah Löber, Björn Rudzewitz, Daniela Verratti Souto, Luisa Ribeiro-Flucht, and Xiaobin Chen. 2024. Developing a Web-Based Intelligent Language Assessment Platform powered by Natural Language Processing Technologies. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 126–136.

Wanjing (Anya) Ma, Michael Flor, and Zuowei Wang. 2025. Automatic generation of inference making questions for reading comprehension assessments. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 398–414, Vienna, Austria. Association for Computational Linguistics.

Michael McGuire and Jenifer Larson-Hall. 2025. Assessing Whisper automatic speech recognition and WER scoring for elicited imitation: Steps toward automation. *Research Methods in Applied Linguistics*, 4(1):100197.

Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741.

Meik Michalke, Earl Brown, Alberto Mirisola, Alexandre Brulet, Laura Hauser, and Maintainer Meik Michalke. 2021. Package 'korpus'.

Lourdes Ortega, Noriko Iwashita, John M Norris, and Sara Rabie. 2002. An investigation of elicited imitation tasks in crosslinguistic SLA research. In *Second Language Research Forum, Toronto*, pages 3–6. Paper presentation.

Anne O'Keeffe and Geraldine Mark. 2017. The English Grammar Profile of learner competence: Methodology and key findings. *International Journal of Corpus Linguistics*, 22(4):457–489.

W James Popham. 2000. The mismeasurement of educational quality. *School Administrator*, 57(11):12–15.

Yves Rosseel. 2012. lavaan: An r package for structural equation modeling. *Journal of statistical software*, 48:1–36.

Nelly Sagirov and Xiaobin Chen. 2025. POLKE: A system for comprehensively annotating pedagogically-oriented grammatical structure use in language production. *Manuscript submitted for publication to Behavior Research Methods*.

Ayfer Sayin and Mark Gierl. 2024. Using OpenAI GPT to generate reading comprehension items. *Educational Measurement: Issues and Practice*, 43(1):5–18.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine Learning–Driven Language Assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Megan Solon, Hae In Park, Carly Henderson, and Marzieh Dehghan-Chaleshtori. 2019. Revisiting the Spanish Elicited Imitation Task: A tool for assessing advanced language learners? *Studies in Second Language Acquisition*, 41(5):1027–1053.

Yishen Song, Junlei Du, and Qinhua Zheng. 2025. Automatic item generation for educational assessments: a systematic literature review. *Interactive Learning Environments*, pages 1–20.

Nina Spada, Julie Li-Ju Shiu, and Yasuyo Tomita. 2015. Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, 65(3):723–751.

Bin Tan, Nour Armoush, Elisabetta Mazzullo, Okan Bulut, and Mark Gierl. 2024. A review of automatic item generation techniques leveraging large language models. *International Journal of Assessment Tools in Education*, 12(2):317–340.

Alessandro Vanzo, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2025. GPT-4 as a homework tutor can improve student engagement and learning outcomes. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31119–31136.

Hadley Wickham. 2019. Package 'stringr'. *http://stringr. tidyverse. org*.

Shu-Ling Wu and Lourdes Ortega. 2013. Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals*, 46(4):680–704.

Shu-Ling Wu, Yee Pin Tio, and Lourdes Ortega. 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.

Xun Yan, Yuyun Lei, and Chilin Shih. 2020. A corpus-driven, curriculum-based chinese elicited imitation test in US universities. *Foreign Language Annals*, 53(4):704–732.

A Appendix A

GPT-generated EIT test items

1. She is buying a new jacket.
2. The blue car is parked outside.
3. The store closes early on Sundays.
4. He usually takes the bus to work.
5. I don't know why she left so suddenly.
6. We had a great time at the beach yesterday.
7. The restaurant downtown serves excellent seafood.
8. She was excited to finally start her new job.
9. The dog that lives next door barks all the time.
10. I hope to visit my grandmother this weekend.
11. The weather in this city changes very quickly.
12. She spent the whole afternoon reading a good book.
13. My neighbor's cat always sits by the window and watches birds.
14. The teacher gave us an interesting assignment about history.
15. The bakery on the corner sells delicious chocolate cake.
16. I wish I could have stayed longer at the party last night.
17. He told me that he would help me move to my new apartment.
18. The little girl with curly hair is playing in the garden.
19. We decided to take a road trip through the mountains next week.
20. If I had known about the meeting, I would have arrived earlier.
21. She was surprised to see her childhood friend after so many years.
22. The scientist explained how the new discovery could change medicine.

23. He was looking for his wallet but couldn't remember where he left it.
24. The old library downtown has a huge collection of rare books.
25. I didn't expect the movie to be so emotional and thought-provoking.
26. The woman who just moved in next door seems very friendly and kind.
27. We were planning to go hiking, but the weather turned out terrible.
28. The students who studied all night found the exam easier than expected.
29. If I had known how difficult this course would be, I might have chosen another.
30. The train that was supposed to arrive at noon has been delayed due to bad weather.

B Appendix B

Complexity measures employed in this study

- Mean word length
- MATTR of lemmas
- Mean words per phrase
- Mean phrases per clause
- Mean clauses per T-Unit
- Normalised rate of occurrence of dependent clauses
- SUBTLEXus log-10 mean frequency of all tokens

C Appendix C

Could you create an Elicited Imitation Test? Please create 30 items, varying in length, start with the shortest one, then gradually generate longer items. Start at 7 syllables.

Below, you will find an example of an existing EIT (in brackets are the amount of syllables of the sentences).

Please only respond with your generated items.
{examples}