International Journal of Language and Literary Studies
Volume 6, Issue 1, 2024
Homepage : http://ijlls.org/index.php/ijlls

Predicting Elicited Imitation Performance from Complexity, Accuracy and Fluency (CAF) of L2 Oral Production

Ghadah Albarqi

Foreign Language Department, Taif University, P.O. Box: 11099, Taif 21944, Saudi Arabia

gadah.g@tu.edu.sa

DOI: https://doi.org/10.36892/ijlls.v6i1.1569

APA Citation: Albarqi, G. (2024). Predicting Elicited Imitation Performance from Complexity, Accuracy and Fluency (CAF) of L2 Oral Production. *International Journal of Language and Literary Studies*, *6*(1), 226–242. https://doi.org/10.36892/ijlls.v6i1.1569

Received: *30/12/2023*

Accepted: *26/02/2024*

Keywords: *Elicited Imitation Test, second language proficiency, complexity, accuracy, fluency Introduction.*

Abstract

The Elicited Imitation Test (EIT) is a testing tool that has been used to measure the second language (L2) proficiency for L2 research. The EIT construct is, nevertheless, still not comprehensively investigated in EFL contexts. This study aims to contribute to this field by examining the extent to which complexity (indexed by lexical diversity, mean length of utterance), accuracy (percentage of error-free clauses) and fluency (breakdown, self-repair, speed) (CAF) can predict EIT performance in a Saudi context. A total of 66 learners were recruited to narrate picture stories. Multiple regression analysis was employed to answer the research questions. The findings show that speed fluency, filled pauses, and accuracy are key indicators of EIT performance. This implies that faster and more accurate L2 speech with fewer pauses is likely associated with a better EIT performance. Additionally, the data highlights the validity of EIT as an assessment instrument across various linguistic settings.

1. INTRODUCTION

The EIT has emerged as a prominent subject of interest in L2 research (Bley–Vroman & Chaudron, 1994; Gaillard & Tremblay, 2016; Kim et al., 2016; Ortega et al., 2002; Suzuki & DeKeyser, 2015; Wu et al., 2022). L2 researchers have utilized EITs, also known as sentence repetition tests, across different languages. For instance, Ortega et al. (2002) developed an EIT tool for English, German, Japanese, and Spanish. Moreover, EITs have been formulated in other languages, such as French (Gaillard & Tremblay, 2016; Tracy-Ventura et al., 2014) and Korean (Kim et al., 2016).

While EITs have been instrumental in assessing L2 proficiency across diverse languages, the specific construct they assess remains under debate (see McDade et al., 1982; Suzuki & DeKeyser, 2015). One promising approach to elucidate the underlying construct of EIT lies in investigating its relationship with Complexity, Accuracy, and Fluency (CAF): specific linguistic features derived from L2 spoken production (see Housen et al., 2012; Housen & Kuiken, 2009). The rationale for employing CAF triad stems from L2 research which suggests that CAF triad offers insights into processing efficiency; a facet closely tied to the seamless articulation of spoken language (Gaillard & Tremblay, 2016; Van Moere, 2012). Hence,

examining the relationship between CAF and EITs can offer insights into the construct that EIT evaluates. Nonetheless, a perusal of extant literature reveals diverse approaches in operationalizing the CAF dimensions, which results in a disparity in outcomes (elaborated in the next section). The current research aims to contribute to the ongoing debate by examining how well the CAF triad can predict EIT performance.

On the other hand, despite the substantial volume of work dedicated to EIT research, it is important to note that studies in this domain have predominantly focused on recruiting participants with L1 English mainly in the United States (Isbell & Son, 2022). Little is known about whether the findings of these studies can be applied to EFL learners in different contexts, such as EFL learners in a Saudi context, whose first language (Arabic) differs typologically from English (Alzahrani & Algethami, 2023). Thus, this paper seeks to address this notable gap in the literature by examining the predictive validity of CAF triad into EIT performance in an EFL context with Arabic as L1 and English as L2. Through this approach, the current study aims to shed light on the universal applicability and validity of the EIT in assessing language proficiency across a spectrum of linguistic contexts.

1.1. Elicited Imitation Test

The EIT has been employed in research contexts to evaluate learners' L2 proficiency in a spoken mode (Bley–Vroman & Chaudron, 1994; Gaillard & Tremblay, 2016; Kim et al., 2016; Ortega et al., 2002; Suzuki & DeKeyser, 2015; Wu et al., 2022). The rationale behind EIT use is rooted in psycholinguistics.

To accurately repeat a sentence, speakers need to break it into grammatical elements and construct a cognitive representation using their linguistic system (Bley-Vroman & Chaudron, 1994). This process is seen as reconstructive, requiring more than just memory-based repetition (Bley-Vroman & Chaudron, 1994). Various studies indicate that EIT can assess different aspects of language, including L2 syntactic knowledge (Ellis, 2005; Erlam, 2006), morphology development (West, 2012), and L2 listening comprehension (Jensen & Vinther, 2003). EITs also offer valuable insights into learners' vocabulary and fluency (Wu & Ortega, 2013). Similarly, L2 research has indicated that EITs assess L2 learners' grammar and their implicit knowledge (Gass, 2018). In essence, if the sentence aligns with an individual's grammatical knowledge, it is likely to be repeated easily, but if it does not, learners may adjust the sentence to match their current grammatical knowledge (Gass, 2018). Additionally, studies have shown that EITs work better when measuring global language constructs, such as overall language proficiency or implicit understanding of grammar, as opposed to specific linguistic knowledge or skills like syntax or phonology (see Hulstijn, 2011; Klem et al., 2015; Yan et al., 2016).

External validity of EITs has been explored through examining their correlations with criterion measures such as CAF dimensions. CAF measures are derived from oral performances elicited during communicative tasks, like the description of narrative-based pictures. It has been suggested that external validity of a test can be established by comparing its scores with assessments of abilities derived from sources beyond the test itself (Alderson et al., 1995; Kane, 2013). In the case of EITs, evidence for their external validity has been promising (Wu et al., 2022). Researchers have found that EITs measure a construct closely related to automatized

language knowledge, aligning with the proficiency assessed in oral interviews and demonstrating shared variance with complexity, accuracy, and fluency measures of spontaneous speech (Isbell & Son, 2022).

However, despite the existing validation research that supports the use of EITs as measures of oral proficiency, there are notable gaps in the current literature. These gaps include the need to verify whether EIT scores can be predicted by CAF triad across various EFL contexts. Previous studies have predominantly centered on speakers with L1 backgrounds such as English, along with learners of various L2s, including German, Japanese, Spanish, French, Korean, and Mandarin. Notably, the examination of L2 English has been comparatively limited in scale when compared to the aforementioned languages (Wu et al., 2022). This challenges the generalizability of EIT scores across diverse EFL contexts. There is a pressing need to assess the validity of EIT scores in novel contexts with learners from L1s that are underrepresented in EIT research such as L1 Arabic. This endeavor holds the potential to enhance the validity of the EIT. As such, this study aims to examine the EIT within a foreign language setting, focusing on learners with L1 Arabic and L2 English.

1.2.Complexity, Accuracy, and Fluency Dimensions

CAF aspects are linguistic features obtained from L2 speech and employed as indices of capturing L2 oral proficiency. The CAF triad is commonly employed as descriptors of L2 oral and written assessment and serves as indicators of L2 processing and development (De Graaff & Housen, 2009; Housen et al., 2012; Housen & Kuiken, 2009; Skehan, 1998, 2003; Vercellotti, 2015).

Skehan (1996, 1998) introduced CAF through his proficiency model for the first time. The definitions and operationalisations of CAF features are still controversial particularly complexity which entails different meanings across L2 studies (Housen et al., 2012; Norris & Ortega, 2009; Skehan, 2003; Tavakoli & Skehan, 2005). Complexity commonly refers to cognitive and linguistic complexity (Housen et al., 2012). Cognitive complexity means processing difficulty which might be the result of varied factors such as the learners' L1 background, task types or task conditions. On the other hand, linguistic complexity pertains to the properties of L2 features, including syntactic complexity and lexical diversity (Housen et al., 2012; Skehan, 1996, 1998). The latter will be the focus of the current investigation.

Unlike the concept of complexity, there is less disagreement on the definition of accuracy. It is defined as any deviation from a norm or native-like use of language including diverse types of errors (grammatical, phonological, and lexical) (Housen et al., 2012; Housen & Kuiken, 2009). Investigating fluency, Tavakoli and Skehan (2005) identified three dimensions: speed fluency (speech rate) breakdown fluency (frequency, length, types, and places of pauses), and repair fluency (hesitations, reformulations, and repetitions). This fluency model has been examined and validated in several studies (e.g., Tavakoli & Skehan, 2005; Skehan et al., 2016; Suzuki & Kormos, 2023). Researchers suggest that all the three CAF dimensions need to be employed together when assessing L2 learners' proficiency and performance (Housen et al., 2012; Ortega 1995; Skehan, 2009; Tavakoli & Skehan, 2005). Operationalisations of CAF features will be explained later in the '*Coding CAF Measures' Section*.

To understand the construct of CAF, researchers such as Skehan (1998), Skehan et al. (2016) and Robinson (1997, 2015) investigated CAF dimensions in relation to the underlying speech production components: conceptualization, formulation, and articulation. Skehan (2009) argues that establishing associations between CAF dimensions and Levelt's (1989) cognitive processing model is useful since it is the most reliable and validated model for describing speech processing. The first component of Levelt's model is the conceptualizer which is responsible for generating ideas (a preverbal message). At the formulator stage, the grammatical encoding is conducted which entails lexical access, syntactic and phonological encoding. The articulator, finally, executes the speech which is revised by the monitor before and after articulation (Levelt, 1989). Unlike L1 speech, attentional resources in L2 are "limited" and "L2 speech processing frequently needs attention at the level of lexical, syntactic and phonological processing" (Kormos, 2006. p.173). Due to the limitation in attentional resources, L2 speech processes may not be automatic (i.e., fast, effortless, attention-free) especially at lower levels of proficiency (see DeKeyser, 2001; Segalowitz, 2003; Tavakoli, 2019). As such, cautions should be exercised when interpreting L2 findings in relation to Levelt's model (See Kormos, 2006; Skehan, 2009).

Research findings suggest that complexity is connected to the conceptualiser (See Ellis, 2008; Pallotti, 2009; Skehan, 2009; Tavakoli & Foster, 2011). Specifically, increasing the pressure at the conceptualiser stage (e.g., increasing task difficulty) leads to the production of more complex language (Skehan, 2009). Accuracy, on the other hand, is more likely to be affected by the influences of the Formulator (Palloti, 2009; Skehan, 2009). When certain processes at the formulator stage become automatised as a result of proficiency development such as syntactic encoding (Robinson, 1997), a more accurate L2 speech will be delivered (Albarqi & Tavakoli, 2022; Nakatsuhara et al., 2019). With regard to fluency, it is a multifaceted construct that can be affected by the influence of these three components, Conceptualiser, Formulator, and Articulator. There is arguably a competitive relationship (trade-off) among these measures (Robinson, 2015; Skehan, 1998; 2003) due to the limited cognitive resources of working memory; that is, effective performance on one dimension (e.g., complexity) would affect the production of the other (e.g., fluency, accuracy). Interpreting the findings of research on the CAF triad in relation to Levelt's model could enhance our understanding of L2 processing and development.

L2 researchers have observed a relationship between EIT scores and CAF measures in spontaneous speech (Isbell & Son, 2022; Kim et al., 2016; McManus & Liu, 2020; Park et al., 2020; Tracy-Ventura et al., 2014). Specifically, Tracy-Ventura et al. (2014) found correlation between EIT and speed fluency (operationalized as syllables per minute) in L2 French. In another study, Kim et al. (2016) noted correlations of EIT with measures such as syntactic complexity (number of morphemes per clause), accuracy (correct clause rate), and fluency (number of morphemes per minute) in L2 Korean. Park et al. (2020) proposed predictive indicators for EIT in L2 Spanish based on complexity (number of words per number of AS-units), accuracy (number of error-free clauses divided by number of clauses), and fluency (length of pauses, syllables divided by total speech time). Additionally, McManus & Liu (2020) revealed an association between EIT for L2 Mandarin and lexical complexity (types of motion verbs), accuracy (motion clauses), and fluency (number of clauses). Notably, these studies

employed varied operationalization of CAF dimensions, leading to inconsistent interpretations of the relationship between EIT and CAF. This study aims to contribute to the existing body of research by examining the extent to which the CAF triad can predict EIT performance, thereby clarifying the underlying construct the test assesses.

1.3.Aims of the study

The academic literature reviewed in the previous sections indicates a relationship between EIT and CAF dimensions. However, the findings have been inconsistent, underscoring the need for further research. To address this gap, the present study aims to determine how well CAF measures can predict EIT performance within the study's context. This study seeks to address the following research questions:

Q1- To what extent can complexity indexed by lexical diversity (type-token-ratio) and syntactic complexity (mean length of utterance) predict EIT performance?

Q2- To what extent can accuracy (the percentage of error-free clauses) predict EIT performance?

Q3- To what extent can fluency (speed, self-repair, breakdown) predict EIT performance?

METHOD

3.1. Design

This study involved multiple regression analysis with EIT as a dependent (continuous) variable and CAF aspects as independent variables.

3.2. Participants

Participants in the study were 66 female undergraduate students, all native Arabic speakers, with ages ranging from 18 to 23. These individuals shared a common background in second language learning, having undergone formal English instruction for a duration of 8-9 years, encompassing both school and university settings. Notably, none of the participants had prior exposure to living in an English-speaking country. Forty of the students were in the first year, while 26 were in the second year of their bachelor's degree. Each participant provided official written consent for their participation in the study.

3.3. Materials

The study relies on two primary data sources. Firstly, EIT scores were derived from learners' repetition of EIT stimuli, and secondly, L2 oral performance was collected from learners' descriptions of narrative picture prompts.

Elicited Imitation Test

To assess the proficiency level of the participants, the EIT created by Wu and Ortega (2013) was used. The EIT comprises nineteen sentences, with syllables ranging from seven to nineteen, spoken by a native English speaker. One opportunity was given to learners for

listening and repeating the sentences. Scores, spanning from 0 to 4 points, were employed to assess the test items. A perfect repetition was given four points. Three points were allocated for the accurate production of content and meaning with some changes in the form, while two points were assigned for modifications in the form or content that altered meaning. One point was granted for producing half of the test item or less, whereas zero points were given for repeating only a word or being unable to repeat any part of the sentence, as detailed in Appendix A. Following previous studies (e.g., Bley–Vroman & Chaudron, 1994; Gaillard & Tremblay, 2016; Wu & Ortega, 2013), and to ensure the reliability of scoring, the author, along with an expert researcher, coded the data. Any disagreements were resolved through discussion. This procedure was commonly followed in EIT research. The reliability of the EIT was assessed with Cronbach's alpha, and high internal consistency, $\alpha = .92$, demonstrating high reliability. This finding is similar to the one reported by Tracy–Ventura et al. (2014). However, caution needs to be taken when interpreting the data due to the small sample of the study.

Tasks

In the present study, the elicitation of L2 learners' speech was conducted using oral narrative picture prompts which 'are regarded as valid pedagogic tasks which are frequently used across different educational settings for teaching, learning, and assessment purposes' (Tavakoli, 2011, p. 73). This indicates that talking about picture-based prompts is similar to everyday speaking activities and likely requires cognitive efforts comparable to those encountered in real-life situations. Additionally, picture stimuli are convenient to implement, and prove effective in prompting spontaneous oral production (Van Hest, 1996). The literature has suggested various aspects of task design that can affect performance, such as element number in a task, task structure, and storyline complexity (see De Jong & Vercelloti, 2016; Garcia-Ponce & Tavakoli, 2022; Tavakoli & Foster, 2011). The picture prompts used in the current study were similar in terms of the number of elements, structure, level of complexity in the storyline, and frequency of vocabulary (see Appendix B).

Procedures

The data was collected as part of a larger project over four weeks. The EIT was administered, and each student was required to say the test item once the recording stopped. Following that, the participants were given a one-minute pre-task planning period to examine the prompts and prepare their utterances. Instructions were introduced in Arabic (the participants' L1). To reduce practice and order effects, the picture-based tasks were counterbalanced.

Coding CAF Measures

As discussed in the literature review, a number of CAF measures have consensually emerged in L2 studies as valid oral indices of L2 proficiency (e.g., De Graaff & Housen, 2009; Ellis & Barkhuizen, 2005; Housen et al., 2012; Housen & Kuiken, 2009; Skehan, 1998, 2003; Tavakoli & Skehan, 2005; Vercellotti, 2015). They informed the choice of indices used in this study (Table 1). A total of 132 speech samples (66 x 2 performances) were collected from the participants' oral tasks. Following the guidelines introduced by Foster et al. (2000), the transcripts were segmented into AS units.

Table 1: *Operationalisation of Complexity, Accuracy, and Fluency aspects*

Dimension	Measures	Definition
Complexity	Type token ratio (Lexical diversity)	The number of different words in a speech sample (Foster & Tavakoli, 2009)
	Mean length of utterance (Syntactic complexity)	Mean number of words per AS unit (Foster & Tavakoli, 2009)
Accuracy	Percentage of error-free clauses	Error-free clauses were calculated manually, followed by dividing the result by the number of clauses, and finally, multiplying it by 100.
Fluency	Syllable per minute (speed)	Number of syllables produced per minute
	Self-repair (repair)	The global number of self-repairs in a speech sample.
	Frequency of silent pauses (breakdown)	This entails dividing the number of silent pauses by the speech duration (in seconds), and multiplying the result by 60.
	Frequency of filled pauses (breakdown)	This was done by dividing the number of filled pauses by the speech duration (in seconds) and multiplying the result by 60.

Speech samples were coded for measures of CAF following Housen's et al. (2012) research in this area. For lexical complexity, VocabProfilers (Cobb, 2017) was used to calculate the type-token ratio as a measure of lexical diversity. The mean length of utterances was calculated manually by having the text pruned (removing all the filled pauses); then the number of words was counted and divided by the number of AS units. Speed fluency, measured as speech rate, was calculated by determining the number of syllables per minute using the Syllable Count program (www. syllablecount.com). This was conducted in an unpruned text, as every single syllable is crucial for this measure. Manual calculation of accuracy involved dividing the error-free clauses by the total number of clauses in the speech and then multiplying the result by 100 (see Table 1). 10% of the data was second rated by an expert researcher and an agreement of 90% was achieved between the raters for self-repair; 87% for the percentage of error-free clauses, 92% for pauses, 92% for AS units. Rater disagreements were resolved through discussion. Preliminary analyses were conducted to verify the absence of violations of normality, linearity, and homoscedasticity assumptions.

4. RESULTS

This study aimed to examine the extent to which CAF measures can predict performance on EIT. Descriptive statistics for CAF measures can be seen in Appendix C. Multiple regressions were conducted, with EIT as the dependent variable, and CAF measures as independent variables.

Examination of regression analysis assumptions indicated that all variables showed normal distribution, absence of multicollinearity, and a linear relationship between the independent and dependent variables. This suggests that the current data met the necessary requirements of

regression analysis. The findings from the multiple regressions are presented in Table 2. To interpret the strength of f^2, Cohen's (1988) guidelines show that an f^2 value of 0.02 suggests a small effect size, 0.15 indicates a medium effect size, and 0.35 represents a large effect size. However, this study adopts the recent proposal introduced by Plonsky and Ghanbar (2018) to evaluate the findings in second language research. Their guidelines interpret the strength of the adjusted R^2 values as follows: values up to .20 are regarded as small, while those exceeding .50 are considered large (Plonsky & Ghanbar, 2018).

Table 2

Multiple regression model predicting EIT from complexity, accuracy, and fluency

Models		B	SE	β	*t*	*P*	R Square	Adjusted R^2	Effect size (Cohen's f^2)
complexity	Type-token ratio	-14.11	16.90	-.07	-.83	.406	.010	-.005	--
	Mean length of utterance	.32	.45	.06	.72	.473			
Accuracy	Error-free-clauses	.31	.05	.46	5.86	.000	.209	.203	0.25
Fluency	Syllable per minute	.19	.03	.51	6.61	.000	.298	.276	0.38
	Silent pauses	-.13	.23	-.04	-.54	.59			
	Filled pauses	-.58	.14	-.33	-4.31	.000			
	Global repair	-.26	.59	-.03	-.44	.664			

As demonstrated in Table 2, the results of the accuracy and fluency models reached the levels of statistical significance. Regarding accuracy, the model reached statistical significance, F (1,13 = 34.38, $p < .000$), explaining 20.9% of the variance in EIT performance (Adjusted R^2 = .203). The fluency model, F (4,13 = 13.46, $p < .000$), reached a statistically significant level, explaining 29.8% of the variance in EIT performance (Adjusted R^2 = .276) which means that EIT performance can be predicted from certain fluency features. Syllables per minute and filled pauses significantly contributed to the model of fluency ($p < .000$). As Table 2 shows, speed fluency (measured by syllable per minute) was the measure that accounted for the greatest variance in the fluency model (51%), followed by filled pauses (33%). Filled pauses showed a negative value which means that producing fewer filled pauses in L2 speech can be associated with a better EIT performance. The complexity measures, on the other hand, did not reach statistical significance in this study (see the Discussion section). Considering Plonsky and

Ghanbar's (2018) guidelines, the findings of adjusted R^2 values are medium. Nevertheless, these findings are meaningful in the context of this study, as discussed in the next section.

Taken together, syllables per minute indicate a unique contribution of 26% (R square multiplied by itself) to the explanation of variance in EIT performance, followed by the percentage of percentage of error-free clauses (22%), and filled pauses (11%). This means that L2 speech which tends to be faster, more accurate, and has fewer filled pauses would probably be associated with a better EIT score. In summary, the present data revealed that accuracy (percentage of error-free clauses) and fluency (syllables per minute and filled pauses) could predict EIT performance whereas the complexity measures could not in the context of the study.

5. DISCUSSION

The primary objective of the present study was to investigate whether CAF aspects could serve as predictors for learners' EIT performance. The analyses indicated that accuracy (measured by the percentage of error-free clauses) and fluency (syllables per minute and filled pauses) significantly predicted L2 learners' EIT performance.

The first research question investigated whether complexity (indexed by type-token ratio and mean length of utterance) can predict EIT performance. The findings indicated that complexity did not predict EIT performance in the present study. This finding was in line with Kim et al. (2016) where no association was found between EIT and complexity in their study. A potential explanation is that L2 learners are likely aware of their limited lexis and work within their limitations (Skehan et al., 2016). Park's et al. (2020) provided a reasonable interpretation for the disparity of their results (i.e., EIT was predicted by lexical complexity). They argued that that lexical complexity might have emerged as a predictor of EIT performance in their study due to their experimental design of the EIT, which discouraged repeating sentences without understanding. In other words, there was a certain level of difficulty in their design which triggered lexical complexity (Park et al., 2020). This interpretation is in line with the theoretical underpinning of the Conceptualiser (see section 3) where increasing task difficulty likely leads to linguistic complexity (Skehan, 2009; Robinson, 2015). This means that tasks used to elicit CAF features can have an impact on the results. For example, Tracy–Ventura et al. (2014) reported that the EIT correlated with lexical diversity when an oral interview was used to elicit oral production; however, it correlated with speech rate in the picture-based tasks. It can be argued that using open tasks such as describing the weekends (as in Park et al., 2020) or an oral interview (in Tracy–Ventura et al., 2014) might encourage learners to use much of their recently acquired L2 lexis which results in producing greater lexical diversity. In short, the trade-off effect produced by task types may likely have a confounding effect on the relationship between the CAF triad and the EIT (for more on the trade-off effect see Garcia-Ponce & Tavakoli, 2022; Robinson, 2015; Skehan & Foster, 1999; Tavakoli & Foster, 2011).

In terms of the non-significant results of syntactic complexity, the current findings diverge from previous research, such as Kim et al. (2016), which identified a correlation between EIT

and syntactic complexity (number of morphemes per clause). Discrepancies in results may arise from varying operationalizations of syntactic complexity and the distinct tasks utilized for speech elicitation. While this study used a picture-based narrative, Kim et al. (2016) employed a personal information task. The latter is generally viewed as less demanding for L2 learners than picture-based tasks (Garcia-Ponce & Tavakoli, 2022; Skehan, 1996). As complexity is tied to restructuring and risk-taking (Skehan, 2014), a less challenging task might encourage greater syntactic complexity in L2 speech. Hence, inconsistency in task type and operational definitions might explain the different results observed in the connection between CAF and EIT.

The data of the second research question, on the other hand, demonstrated that accuracy measured by the percentage of error-free clauses significantly predicted EIT performance and accounted for 22% of the variance in EIT performance. This is in line with the findings of previous studies by Park et al. (2022) and Kim et al. (2016). Research suggests that accuracy (indexed by the percentage of error-free clauses) is closely linked to a higher proficiency level (Albarqi & Tavakoli, 2022; Nakatsuhara et al., 2019; Tavakoli et al., 2016). This means that producing a higher percentage of error-free clauses suggests that L2 elements are restructured and adjusted because of the development in the L2 interlanguage system (Housen & Kuiken, 2009). This finding is important in the context of the study because it shows that the ability to repeat sentences in EIT test may largely draw on L2 learners' interlanguage system. This measure has been criticised, for not assessing the gravity of the error (Foster & Wigglesworth, 2016). Nevertheless, this measure is regarded as the most reliable and straightforward measure in Task Based Language Teaching (TBLT) research (Park et al., 2022; Skehan, 2009; Skehan et al., 2021; Tavakoli, 2019). Additionally, mounting evidence supports its effectiveness as an indicator of language proficiency (Albarqi & Tavakoli, 2022; Nakatsuhara et al., 2019; Park et al., 2022; Tavakoli et al., 2016). As such, adopting this measure can promote generalizability across L2 studies.

The third research question examined the contribution of fluency features (speed, breakdown, repair) to EIT performance. The results indicated that speech rate (measured by syllables per minute) emerged as the most prominent predictor of EIT performance in the present study as it contributed 26% to the explanation of variance in EIT performance. This finding is similar to the results of previous studies that employed this measure, such as Kim et al. (2016); Tracy–Ventura et al. (2014) and Wu and Ortega (2013). These studies employed a picture-based narrative which suggests that task type might have a positive impact on speech rate. The analysis also indicated that filled pauses emerged as a significant predictor of EIT performance, accounting for 11% of variations in L2 learners' performance on EIT. This means that the ability to speak faster with fewer filled pauses is likely associated with a better EIT performance. A possible explanation is that with proficiency development, some production processes at the Formulator stage may become automatised (see DeKeyser, 2001; Kormos, 2006; Segalowitz, 2003; Tavakoli, 2019). Other studies also reported that lexical and syntactic encodings can reach automatisation as a result of proficiency development (Pellicer-Sánchez, 2015; Robinson, 1997). The automatisation process, characterised by attributes like speed, effortlessness, and reduced attentional demand, can lead to faster speech production (DeKeyser, 2001; Segalowitz, 2003; Tavakoli, 2019); fewer filled pauses (Albarqi & Tavakoli,

2022); and different types of self-repair (Kormos, 2000). However, research found that fluency features could be the result of either personal speaking style (De Jong et al., 2015; Derwing et al., 2009) or L1 influence (Duran-Karaoz & Tavakoli, 2020). As such, it is advised to exercise caution when analysing and interpreting fluency data. Regarding the non-significant findings related to silent pauses and repairs, this study utilized a global measure for both, which might not have been adequately sensitive to predict EIT performance. For a more nuanced comprehension of these fluency aspects, future studies should consider adopting specific measures for silent pauses, like calculating silent pauses within clause boundaries and at the clause boundaries (see Tavakoli, 2011), as well as distinct measures for repairs, including hesitations, repetitions, and false starts. Such specific measures could potentially offer insights into the associations with EIT performance.

In short, by indicating that EIT performance is best predicted by fluency (speech rate and filled pauses) and accuracy (percentage of error-free clauses), the present study confirms the reconstructive nature of the EIT. The current findings support the existing research findings that EIT can serve as an accurate indicator of changes in an L2 learner's interlanguage system and that it can be used as a valid proficiency assessment tool in L2 research.

6. CONCLUSION

The purpose of this study was to investigate the predictive variables for EIT performance by employing CAF dimensions. The results presented in this study adds credence to the claim that EIT can assess L2 learners' linguistic ability indicating that accuracy and fluency dimensions are the most prominent predictors of EIT performance. That is, sentence repetition draws upon the L2 interlanguage system (Kim et al., 2016; Okura & Lonsdale, 2012; Park et al., 2020). It is, nonetheless, important to be cautious when interpreting the results, particularly about CAF dimensions since various factors, such as task type and differences in CAF operationalisation, may contribute to variations in learners' performance.

The trade-off effect produced by task types may likely trigger varying effects on the relationship between the CAF triad and the EIT. L2 research indicates that L2 learners tend to produce a higher rate of complex utterances during cognitively demanding tasks while making greater accuracy and/or fluency in simpler tasks (for more on the trade-off effect see Garcia-Ponce & Tavakoli, 2022; Robinson, 2015; Skehan & Foster, 1999; Tavakoli & Foster, 2011). That is, the association between CAF and EIT might be confounded by tasks which have various levels of difficulty. It should be noted that L2 fluency may be affected by speakers' own styles (see Derwing et al., 2009; Duran-Karaoz & Tavakoli, 2020; Skehan & Foster, 2005). For instance, Duran-Karaoz and Tavakoli (2020) found that certain L2 fluency features (i.e., pauses and self-repair), reflect the learners' L1 speaking styles to a considerable degree. This implies that the CAF features elicited by speaking tasks do not necessarily fully reflect learners' proficiency; they may represent learners' speaking styles or their L1 background, and this consequently can affect the interpretation and generalisation of the findings. Hence, it is important for future research to control for L1 styles when examining CAF dimensions.

Future studies also need to explore other predictive variables of the EIT performance such as the role of working memory on specific EIT items, including longer items or sentences with

complex structures. This can help enhance the validity of this test. In summary, the results of this study affirm the validity of EIT and its applicability in the present EFL context. It can, therefore, be used effectively for both diagnostic purposes in research and placement in language classes.

REFERENCES

Albarqi, G. & Tavakoli, P. (2022). The effects of proficiency level and dual task condition on L2 self-monitoring behaviour. *Studies in Second Language Acquisition*, 1-22.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Alzahrani, A., & Algethami, G. (2023). Phonological Awareness and Word Reading Fluency Among Young Saudi Learners of English. *International Journal of Language and Literary Studies,* 5(1), 14-27.

Bley-Vroman, R., & Chaudron, C. (1994). Elicited imitation as a measure of second- language competence. *Research methodology in second-language acquisition*, 245- 261.

Cobb, T. (2017). *Web Vocabprofile*. An adaptation of Heatley, Nation & Coxhead's (2002) Range, http://www.lextutor.ca/vp/

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed). Mahwah: Lawrence Erlbaum Associates.

De Graaff, R., & Housen, A. (2009). Investigating the Effects and Effectiveness of L2 Instruction. In M. Long, & C. Doughty (Eds.). *The handbook of language teaching* (pp. 726-755). Oxford: Blackwell Publishing.

DeKeyser, R. (2001). Automaticity and automatization. In P. Robinson (Ed.) *Cognition and second language instruction* (pp. 125-151). Oxford: Oxford University Press.

De Jong, N.H., Groenhout, R., Schoonen, R.,& Hulstijn, J.H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*, 223-243.

De Jong, N., & Vercelloti, M. L. (2016). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, *20*(3), 387-404.

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1fluency and L2 fluency development. *Studies in Second Language Acquisition*, 31, 533–557.

Duran-Karaoz, Z., & Tavakoli, P. (2020). Predicting L2 fluency from L1 fluency behavior: The case of L1 Turkish and L2 English speakers. *Studies in Second Language Acquisition*, *42*(4), 671-695.

Ellis, R. (2005). *Planning and task performance in a second language* (Vol. 11). Amsterdam: John Benjamins Publishing.

Ellis, R. (2008).*The study of second language acquisition* (2nd ed.). Oxford: Oxford University Press.

Ellis, R. and G. Barkhuizen. (2005). *Analysing Learner Language*. Oxford University Press.

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

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language learning*, 59(4), 866-896.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354-375.

Foster, P. & Wigglesworth, G. (2016). Capturing accuracy in second language performance: the case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*,98-116.

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language learning*, *66*(2), 419-447.
Garcia-Ponce, E. E., & Tavakoli, P. (2022). Effects of task type and language proficiency on dialogic performance and task engagement. *System*, 105.
Gass, S. (2018). SLA elicitation tasks. *The Palgrave handbook of applied linguistics research methodology*, 313-337. Palgrave Macmillan.
Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473.
Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing.
Hulstijn, J. H. (2011). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language Cognition*, 15(2), 422-433.
Isbell, D., & Son, Y. (2022). Measurement Properties of A Standardized Elicited Imitation Test: An Integrative Data Analysis. *Studies In Second Language Acquisition*, *44*(3), 859-885.
Jensen, E. D., & Vinther, T. (2003). Exact repetition as input enhancement in second language acquisition. *Language learning*, 53(3), 373-428.
Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1–73.
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. *Modern Language Journal*, 100, 655–673.
Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C.(2015). Sentence repetition is a measure of children's language skills rather than working memory limitations. *Developmental science*, *18*(1), 146-154.
Kormos, J. (2000). The role of attention in monitoring second language speech production. *Language learning*, *50*(2), 343-384.
Kormos, J. (2006). *Speech production and second language acquisition*. New York: Routledge.
Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
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.
McManus, K., & Liu, Y. (2020). Using elicited imitation to measure global oral proficiency in SLA research: Aclose replication study. *Language Teaching*. *55*(1), 116-135.
Nakatsuhara, F., Tavakoli, P., & Awwad, A. (2019). *Towards a model of multi-dimensional performance of C1 level speakers assessed in the Aptis Speaking Test*. British Council.
Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF ininstructed SLA: The case of complexity. *Applied Linguistics*, *30*(4) ,555-578.
Okura, E., & Lonsdale, D. (2012). Working memory's meager involvement in sentencerepetition tests. In N. Miyake, D. Peebles, &R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 2132–2137). Austin, TX: Cognitive Science Society.
Ortega, L. (1995). The effect of planning in L2 Spanish oral narratives. *Studies in second Language acquisition*, *21*, 108-148.
Ortega, L., Iwashita, N., Norris, J., & Rabie, S. (2002). An investigation of elicited imitation tasks in crosslinguistic SLA research. *Paper presented at the Second Language Research Forum*, Toronto, Canada.
Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied linguistics*, *30*(4), 590-601.

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.

Pellicer-Sánchez, A. (2015). Developing automaticity and speed of lexical access: The effects of incidental and explicit teaching approaches. *Journal of Spanish Language Teaching*, *2*(2), 126-139.

Plonsky, L., & Ghanbar, H. (2018), Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, 102, 713–731.

Robinson, P. (1997). Generalizability and Automaticity of Second Language Learning under Implicit, Incidental, Enhanced, and Instructed Conditions. *Studies In Second Language Acquisition*, *19*(2), 223-247.

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 87-121). Amsterdam: John Benjamins.

Segalowitz, N. (2003). Automaticity and second languages. In C.Doughty, & M.Long (Eds.), *The handbook of second language acquisition* (pp. 383-406). London: Blackwell.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*(1), 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2003) Task-based instruction. *Language Teaching*, *36*(1), 1-14

Skehan, P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*, 510–532.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93-120.

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193-216). John Benjamins.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching*, *54*(2),97-111.

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language learning*, *65*(4), 860-895.

Suzuki, S., & Kormos, J. (2023). The multidimensionality of second language oral fluency: Interfacing cognitive fluency and utterance fluency. *Studies in Second Language Acquisition*, 45(1), 38-64.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT journal*, *65*(1), 71-79.

Tavakoli, P. (2019) Automaticity, fluency and second language task performance. In: Wen, Z. E. and Ahmadian, M. J. (eds.) *Researching L2 Task Performance and Pedagogy*. John Benjamins, Amsterdam, pp. 39-52.

Tavakoli, P., Campbell C., & McCormack J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly 50(2)*, 447–471.

Tavakoli, P., & Foster, P. (2011). Task design and second language performance: The effect of narrative type on learner output. *Language Learning, 61*(1), 37-72.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (Vol. 11, pp. 239-273). Amsterdam: John Benjamins.

Tracy–Ventura, N., McManus, K., Norris, J. M., & Ortega, L. (2014). "Repeat as much as you can": Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 143–166). Clevedon, UK: Multilin- gual Matters.

Van Hest, E. (1996*). Self-repair in L1 and L2 production*. Tilburg: Tilburg University Press.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing*, 29(3), 325-344.

Vercellotti, M. L. (2015). The Development of complexity, accuracy, and fluency in second language performance: Alongitudinal study. *Applied Linguistics*, 1-23.

West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition*, 4(3), 203-222.

Wu, S. L., & 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.

Wu, S. L., 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.

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*, 497–528.

Appendices

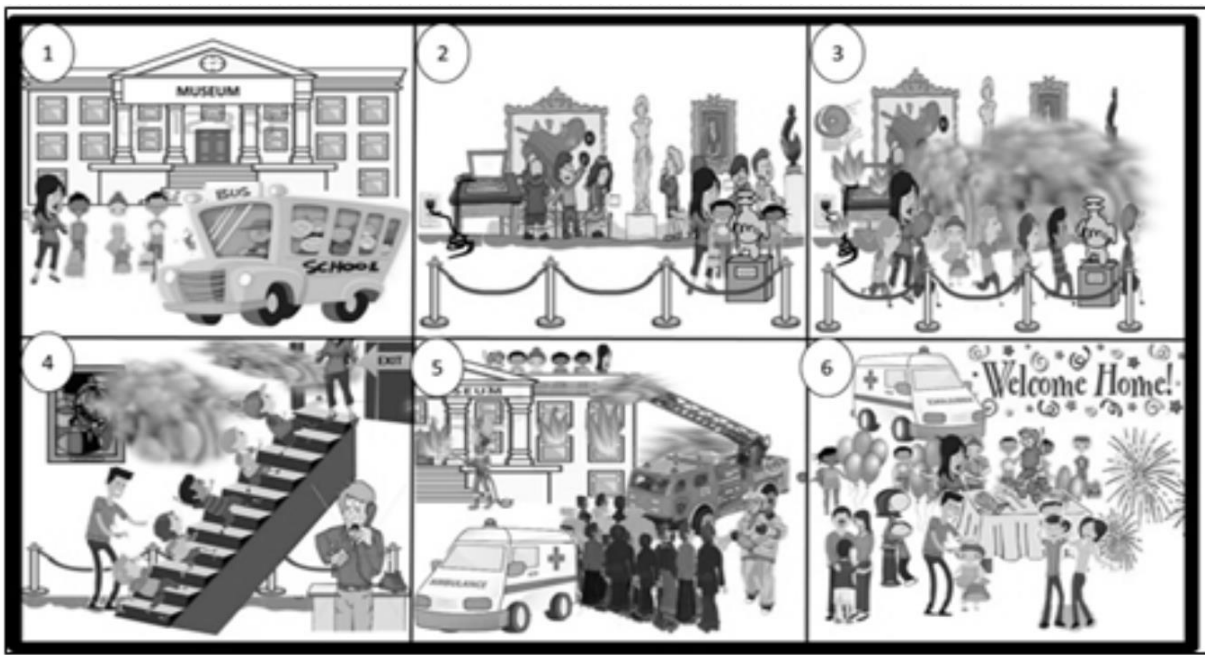
Appendix A: *EIT scoring rubric (based on Ortega et al., 2002)*

Item Score	**Description**	**Examples**
4	Perfect repetition	- The little boy whose kitten died yesterday is sad.
3	Accurate content repetition with some changes of form	-The little boy whose kitten died yesterday feels sad.
2	Changes in content or in form that affect meaning	-The little boy who has kitten died yesterday.
1	Repetition of half or less of the stimulus leading to substantial loss of meaning	- The little boy whose kitten
0	Silence, only one word repeated, or unintelligible repetition	-No response -The boy

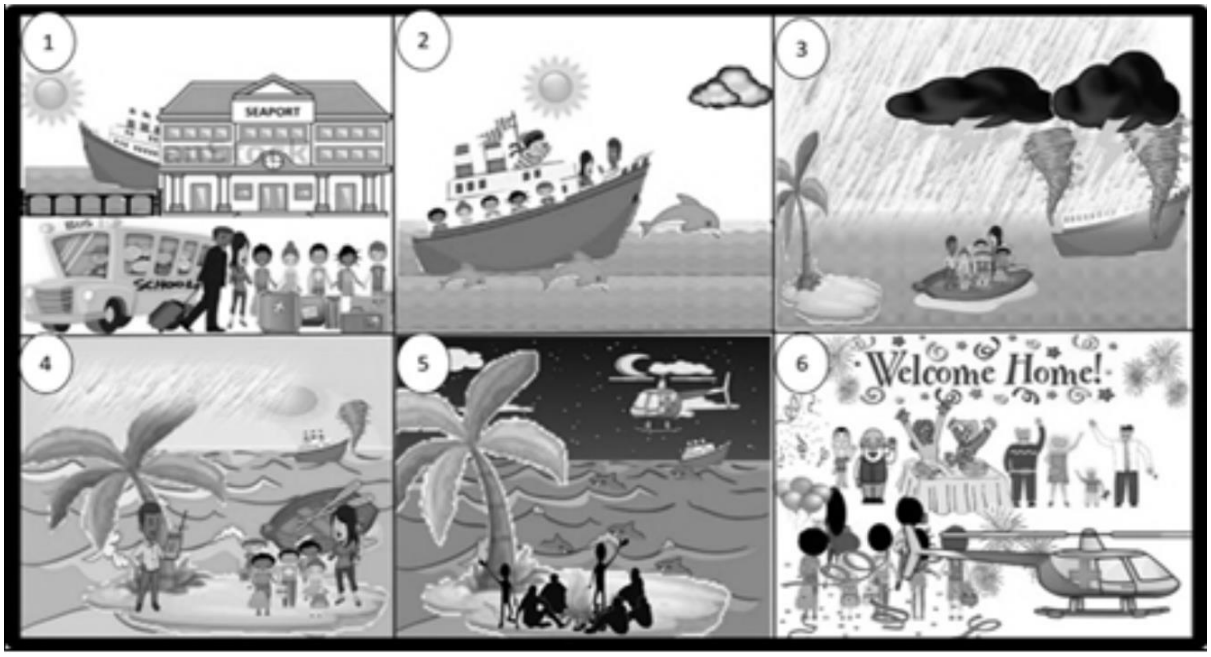
Note. Examples are taken from the data in the current study.

Appendix B: *Picture prompts*

Appendix B-1: Museum



Appendix B-2: Ship



Appendix C: *Descriptive statistics for CAF measures*

Variables	**M *(SD)***	**Min-Max**
Type token ratio	.51(*.07*)	.37- .67
Mean Length of utterance	10.41(*2.56*)	5.0.7- 21
Percentage of error-free clauses	39.49(*19.32*)	.00- 93.7
Syllable per minute (Speed)	126.28(*35.7*)	.73- 212
Frequency of silent pauses (Breakdown)	23.53(*4.21*)	11.0- 34.0
Frequency of filled pauses (Breakdown)	17.54(*7.46*)	2.0- 45.0
Self-repair	2.7(*1.68*)	.00- 7.0

AUTHOR'S BIO

Ghadah ***holds a Ph.D. in Applied Linguistics from the University of Reading. She is a Fellow of Higher Education Academy. Her primary areas of interest encompass psycholinguistics, second language processing, production, and assessment, with a specific focus on self-monitoring. She is also engaged in research related to task-based language teaching (TBLT) and technology-mediated TBLT.***