International Journal of Research in Education Humanities and Commerce

Volume 05, Issue 01 "January - February 2024"

ISSN 2583-0333

ASSESSING EFL SPEAKING BASED ON RECONSTRUCTED CAF MEASURES

Dr. REEM FAHAD ALSHALAN
College of Language Sciences, King Saud University,
Riyadh, Saudi Arabia

https://doi.org/10.37602/IJREHC.2024.5102

ABSTRACT

In applied linguistics research, complexity, accuracy, and fluency (CAF) were based on the importance of task-based activities, task familiarity, communicative competence, or sociolinguistic factors. However, investigating CAF as an independent variable to assess language performance is controversial in research. Some studies found that language cannot be assessed in these dimensions alone since language is complex, changing, and dependent on many factors. Further research shed light on the importance of reconstructing these variables. The purpose aims to test the validity of a speaking assessment scheme based on previous research suggestions to reconstruct CAF measures by comparing it to the TOEFL speaking rubric. There were no statistical differences in the results of the participants in both assessment schemes. The implications are to further design valid assessment schemes based on reconstructed CAF measures, using an automated application that can transcribe and compute these measures instead of calculating them manually.

Keywords: Assessment Schemes, CAF Measures, EFL Speaking, Language Performance.

1.0 INTRODUCTION

Complexity, accuracy, and fluency (CAF) serves as the most common constructs of applied linguistics research. These constructs are applied as either dependent or independent variables. Mostly, second language acquisition researchers use CAF as dependent variables, as it is helpful in assessing the outcomes and theories, that were based on the importance of task-based activities (Ellis, 2012; Robinson, 2001; Skehan, 2009), task familiarity (Sample & Michel, 2014) communicative competence (Pallotti, 2009), or sociolinguistic factors (Larsen-Freeman, 2009).

Contrary to this, investigations related to CAF and its use as an independent variable to assess the language performance is controversial in research (Housen et al., 2012). Larsen-Freeman (2009), indicated that language cannot be assessed in these dimensions alone, as it is complex, dynamic and dependent on many factors (Larsen-Freeman, 2009). As for instance; Housen & Kuiken (2009), mentioned that CAF have been examined and assessed across several different languages. The assessment was held by using different tools, such as subjective and holistic ratings of individual experts that are effective in underlying the more definite levels of ESL proficiency among learners including all the skills. Further research has shed light on the importance of reconstructing these variables to be able to assess the outcomes in a significant way. For example, Skehan (2009) discussed the importance of adding Lexicon to CAF, Pallooti (2009) and Housen et al. (2012) mentioned that communicative understanding must be added and treated separately from CAF measures. Larsen-Freeman (2009), on the other hand shed light on considering sociolinguistic factors when assessing performance through CAF.

Therefore, the purpose of this paper is categorized under two-folds. Firstly, the researcher intends to shed light on the concept of reconstructing CAF and its significance in assessing second language performance. This part of the research will be achieved by focusing on the following characteristics;

- The identification of CAF in research
- CAF operations in research
- Limitations of CAF in research
- Implications of SLA in research and its role in minimizing CAF limitations

Second part of the discussion works with an aim to test the speaking assessment scheme depending on suggestions provided by previous researchers that are helpful in reconstructing CAF measures. The overall process will be held through a comparative analysis between speaking assessment scheme and TOEFL speaking rubric (ETS, 2014). Discussions and analysis related to the statistical significance and differences in score means are also mentioned.

Therefore, research question that is intended to be addressed in this study can be given as;

- Is there a significance difference in the total scores of the participants on the CAF scheme compared to their scores on the TOEFL rubric?”

1.1 Defining CAF in SLA Research

CAF has been widely discussed in research to provide better understanding regarding the fundamental constructs of CAF along with their ability to reflect the oral and written performances of second language learners, language proficiency, along with the progress and development in language learning (Housen & Kuiken, 2009). However, diverse definitions and measures of CAF can be noted throughout the literature since the first attempt to provide major differentiation between language fluency and accuracy (Housen & Kuiken, 2009).

According to Housen and Kuiken et al. (2009), Skehan (1998), was the first to set the notions of CAF as valid tools that can be used to assess the language performance. The discussion regarding the CAF was first mentioned in dual mode hypothesis, which is based on rules and exemplars for input, central and output processing, followed by the trade-off effect of language processing. Previous research findings such as the input theory, the output hypothesis, and negotiation of meaning were challenged, as a significant emphasis was provided to the idea that performance analyses can be developed by involving some of the CAF constructs, rather than utilizing all, such as; concentrating on improving accuracy and complexity over fluency. It was further argued that this trade-off effect could be solved by following the dual mode hypothesis in relying on rules and exemplars to develop SLA performance. Suggestions regarding the importance of task-based activities in developing all notions of CAF without trade-off effects were also provided (Skehan, 1998).

Following CAF notions as an evidence to justify the cognitive theory in SLA, Skehan et al. (1998) highlighted the importance of these constructs in assessing SLA performance. Since then, researchers in the 90s used CAF as a valid assessment tool to measure performance (Housen & Kuiken, 2009).

The most general definition of CAF that Skehan (1998), highlighted stated complexity as the degree of a complicated structure that is varied and elaborated. Accuracy was further defined as an error-free articulation. He addressed fluency as the ability to articulate the language with ease in terms of pace, ideas and the natural flow of the language that is similar to native-like language.

The definition of CAF in Skehan (1998), had its importance in understanding the general features of CAF and what they reflected, however, a deeper understanding of the nature of these constructs and their dimensions was deemed crucial prior to assessing them (Larsen-Freeman, 2009). Looking at complexity, for example, as a notion of expressing varied and elaborated structure was logical, but it did not identify the dimensions of this construct because what was considered varied and elaborated should have been identified in detail (Palloti, 2009). In addition, looking at Accuracy as an error-free articulation was also broad and general because uttering correct sentences may not be accurate in a certain context (Larson-freeman, 2009). As for fluency, articulating with "ease" may be a result beyond language use. Many factors can dominate one's pace, ideas and the natural flow of the language (Larson-freeman, 2009).

The definitions of the notions of CAF may seem simple and straight forward; however, only when these variables are operationalized to assess performance, their complexity and limitations truly be noticed. The following will shed light on how literature measured CAF in SLA research and what were the limitations declared.

1.2 Measuring CAF in SLA research

Defining CAF may seem obvious at first to language instructors, teachers and researchers in applied linguistics; however, examining these constructs in more detail is a complex task which has led to many controversies and ambiguity, especially when trying to operationalize these constructs as variables and measuring the output based on them (Skehan, 2009). The three major analyzing tools which have been used to measure CAF throughout the literature were T-unit analysis, AS unit analysis, and idea unit analysis.

To operationalize CAF, Larson Freeman (2006) assessed both oral and written performance through CAF by utilizing T- units in the assessment. The concept of T-units has been defined as "one main clause with all subordinate clauses attached to it" (Hunt, 1965). Freeman (2006), measured complexity by calculating the total number of clauses divided by the total number of T-units. Accuracy on the other hand was measured through the proportion of error-free T-units to total T-units (in terms of lexical, morphological, and syntactic errors), and measured fluency by the average number of words per T-unit.

Moreover, Foster & Skehan (1996), addressed the common methods of operationalizing CAF in literature which did not apply T-units, but instead followed Foster, Tonkyn & Wigglesworth (2000), modified AS-unit which was defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (p.365). In those studies, Complexity was measured by dividing the total number of clauses by the number of Analysis of Speech Units (AS units). Accuracy was measured through error-free clauses, and fluency was measured in terms of Breakdown Fluency, the number of pauses and the total amount of silence; and also, in terms of Repair Fluency, like reformulation, repetition, false starts etc. (Foster & Skehan, 1996).

Later, Ellis (2012), set valid and detailed measures of CAF, and addressed an idea-unit analysis which valued the meaning and semantic analysis of units. Ellis and Barkhuizen (2005), defined the idea unit as "a message segment consisting of a topic and comment that is separated from contiguous units syntactically and/or intonationally" (p.154) The method of calculating fluency involved removing dysfluencies and counting the number of syllables, pauses, and repetitions and dividing syllables or pauses by total time of speaking or the number of pauses. Accuracy was calculated by percentages taken from calculating the number of error-free clauses and analyzing correct responses regarding certain grammatical features. In addition, Complexity was measured by calculating the number of subordinate clauses in comparison to the other used clauses. Complexity also involved calculating the percentage of the lexical variety of the words produced (Ellis, 2012).

Since AS unit of assessment has modified the T-unit assessment regarding spoken performance, it will be considered in this study. Furthermore, idea-unit assessment will also be used to include one syntactic unit (AS- unit) and one semantic unit (idea- unit) of assessment.

1.3 Limitations of measuring CAF in SLA research

The common measures of CAF focused on the general features of these constructs. They familiarized novice readers with the concept of CAF and displayed a broader explanation of their characteristics as assessment tools. However, when researchers attempt to investigate their theories and generalize their hypothesis based on CAF as valid assessment tools, it is crucial that they have a deeper understanding of these constructs to be able to assess language performance validly (De Jong & Vercellotti, 2016).

Pallotti (2009), challenged CAF measures by arguing that some aspects of CAF are effective in showing the higher performance in quantitative measures but not necessarily reflect improvement in those constructs. In addition, Palloti argued for the notion of complexity, according to which elaborated language and varied lexicon should not be used to measure complexity because they are the features that help to develop it. Pallotti also challenged fluency measures by clarifying that many types of fluency were not taken into consideration when measuring it by the number of pauses and repetition. He further shed light on the importance of adequacy in achieving communication goals and argued that assessing language performance through accuracy can be error free and grammatically correct but may not be understood communicatively. Calculating error-free utterances without considering their communicative role will show unreliable analysis (Palloti, 2009).

Foster & Skehan (1996) acknowledged the need to reconstruct complexity measures, while challenging the accuracy and fluency measures of language in their study. They argued that when measuring accuracy, there was a possible inflation in scores if a speaker's production consisted of many short error-free utterances. Thus, it would negatively affect reporting reliable analytical statistics. For measuring fluency, they argued that the number of pauses and total silence was a major limitation, since native speakers may also pause while speaking. This limitation of fluency showed that calculating an aspect based on general features can cause invalid and unreliable assessments. The fact that the number of pauses were considered and calculated as errors in language performance, while native speakers act in the same way, was a clear indicator of the importance of investigating the concept of fluency of a native speaker before assessing that of the second language performance. This is due to the idea that native speakers also pause while speaking. Furthermore, Skehan (2009) stated that CAF notions did not emphasize the importance of assessing lexical performance despite its large impact on developing L2 proficiency. Lexical development has only been addressed under accuracy (Skehan, 2009) and under complexity (Ellis, 2012), but not as a separate construct.

One of the major limitations of CAF is that they do not include social factors such as context which are essential since language is a social phenomenon (Norris & Ortega, 2006). These limitations of CAF in literature are due to the complexity of the process of second language acquisition and the complexity of language in general, which can be solved by broadening the mind. Conducting further researches in other domains that can capture a valid assessment of performance is of fundamental importance (Larsen-Freeman, 2009). Therefore, the research started investigating other aspects that could decrease the limitations of CA

1.4 Suggestions in SLA research to minimize limitations of CAF

Investigating CAF in SLA research caused a rise in the need to minimize the gap between theory and measurements (Larson-Freeman, 2009). The following will discuss research suggestions to minimize the limitations of CAF declared previously.

Vercellotti (2012), who conducted a longitudinal study on CAF found that they lacked individual paths in development and lacked trade-off effects between them. He acknowledged the importance of CAF measures in assessing performance. However, Vercellotti research supported the importance of choosing activities that can affect performance development in all aspects of CAF. Task based activities have supported minimizing tradeoff effects between CAF measures (Skehan, 2009). Considering this, the present study adapted a task-based activity from Elllis (2012), which will be used for data collection.

Moreover, some studies suggested solutions for the limitations of accuracy by operationalizing it to the involve notions of phonology, morphology, and lexis instead of native-like word choice alone (Wulff & Gries, 2011; Skehan et al. 2009), who was the first to discuss the concept of CAF, recognized the limitation in assessing lexical performance under accuracy. In his study, it was argued that lexicon was an essential part of assessing performance and that it should be identified as a separate construct in addition to CAF (Skehan, 2009). It was further declared that Lexis held distinguishing features and a demanding role in language development, and the assessing lexicon should reflect the extent of lexical variety to which a speaker uses (Skehan, 2009).

Other valid detailed measures of accuracy and fluency were proposed in Foster and Skehan (1996) that includes reconstructing the assessing accuracy through error-free clauses by ranking them according to their length and the proportion of each accurate word length which is computed to prevent score inflation caused by short error-free clauses. Foster and Skehan (1996), reconstructed fluency measures by considering the pauses at the end of clause and mid clause, following native like characteristics where the pause is most frequent at the end and not in the middle of a clause. Despite acknowledging the need to reconstruct complexity, Foster and Skehan (1996), did not address it in their paper.

Pallotti (2009), worked as a prominent figure here and challenged the measures of complexity as mentioned in the limitations. He suggested to consider language development as a separate construct since it is influenced by CAF measures. In addition, to minimize limitations of fluency, identification of the type of fluency was considered before measuring it. Another limitation addressed by Pallotti was the importance of communicative adequacy in assessing performance which is an important measure of CAF dismiss.

Purpura (2017), on the other hand identified Pragmatic knowledge as “the mental structures underlying the ability to communicate functional and other implicational meanings” (p.53). He emphasized the importance of considering the meaning conveyed as an important component of assessing language knowledge explicitly and implicitly. He identified seven types of pragmatic meanings, one of which is rhetorical meanings that consider coherence. Coherence will be measured in this study to measure communicative adequacy explicitly. In addition, communicative adequacy will also be considered implicitly throughout the other components.

Considering the importance of CAF measures in assessing performance, and the need to fill in the gap in the literature by implementing previous research suggestions to decrease the limitations of CAF, this paper attempted to test speaking assessment scheme based on previous suggestions in research. It tested it for reliability and validity following Allen & Knight (2009) and Mackey & Gass (2016).

The scheme of this study was compared to the TOEFL speaking rubric (ETS, 2014) to investigate if there were differences in the participants’ results when they were assessed in both measures. The TOEFL speaking rubric (ETS, 2014), which has been subjected to reliability and validity measures, has been chosen due to its general compatibility with the current scheme. That is because it was based on a communicative approach to language learning through different notions of CAF (Jamieson & Poonpon, 2013). CAF features were considered in the TOEFL speaking rubric under three categories which were delivery, language use, and topic development (ETS, 2014). This paper investigated if there were statistically significant differences in the mean results of the participants in the speaking assessment scheme, which is based on reconstructed CAF measures (Henceforth, CAF scheme), in comparison to their results in the TOEFL speaking rubric (ETS, 2014) (Henceforth TOEFL rubric).

Therefore, based on the literature review, following hypothesis will be tested:

H1: There was no statistically significant difference in the means of the total scores of the participants on the CAF scheme compared to their results on the TOEFL rubric

2.0 METHODOLOGY

2.1 Unit of Analysis

The three major units that were used to measure CAF throughout the literature were the t-unit (Hunt 1965), the AS unit (Foster et al., 2000), and the idea unit (Ellis, 2005). The definition of each unit is presented in Table 1. The units of assessment which were considered in this study were a combination of a syntactic unit (AS-unit) following Foster et al. (2000), and a semantic unit (the idea-unit) following Ellis & Barkhuizen (2005). The t-unit that Hunt (1965) established were not considered, since the AS-unit was a modified version of the t-unit. The AS-unit and the idea unit considered in this study were calculated depending on the construct measured. This combination was due to previous literature suggestions in reconstructing CAF as explained in the following section under instruments.

Table 1. Units of Assessment used to measure CAF in literature

Unit of Assessment	Definition
T-unit	“one main clause with all subordinate clauses attached to it” (Hunt 1965, p. 20).
AS-unit	“a single speaker’s utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either” (Foster, Tonkyn and Wigglesworth 2000: 365).
Idea-unit	“a message segment consisting of a topic and comment that is separated from contiguous units syntactically and/or intonationally” (Ellis & Barkhuizen 2005: 154).

2.2 Instruments

This study used three instruments. First includes the task-based activity provided by Ellis (2012), to collect the responses of the participants. The CAF scheme was used to assess the performance of the participants. The TOEFL rubric was also used to assess the performance of the participants. The following will explain every instrument in detail.

2.3 The task-based activity

A task-based activity was used to test the performance of the participants in speaking. It was used to minimize tradeoff effects of one component of CAF over the other (Skehan, 2009; Ellis, 2012). The activity was adapted from Ellis (2012), because it has been validated and tested for reliability (See Appendix 1). According to Ellis (2012), a task should focus on meaning with an information ‘gap’ which is filled by the learner’s linguistic and nonlinguistic knowledge. It was emphasized that a goal of a task should not be the language used, but the use of language as a mean to complete the task (Ellis, 2012). The present study followed this type of activity, as both instruments were used to assess the performance of the participants based on communicative language learning.

Following this, the task-based activity chosen for this study required learners to answer the activity in pairs. They focused on meaning to fill in an information gap task. The task was called “What can you buy”. Both participants were provided with different lists. The first one had to figure out which items on the list can be bought. The second participant had to figure out which items requested by the first participant were not in stock (Ellis, 2012).

2.4 The CAF schemes

The instruments used to assess the speaking performance of the participants in this study were based on previous research suggestions to reconstruct CAF measures (Ellis, 2005; Pallotti, 2009; Skehan, 2009). The scheme was designed in a rubric form. It consisted of five major aspects which were complexity, accuracy, fluency, lexis and communicative adequacy. The underlying language acquisition theory of the scheme, the measurement of each construct, and the scoring criteria have been explained as follows.

The underlying theory. It is noticed that some of the suggestions for the limitations of CAF addressed in the literature were dependent on the linguistic theory of language acquisition that was underlined in these studies. Considering such diverse concepts as a whole would hinder the possibility of reconstructing CAF in one scheme. Therefore, this assessment scheme was based primarily on a communicative learning approach, placing importance on communicative adequacy in interpreting CAF. It has been suggested by Pallotti (2009), as a possible solution that would increase the validity and reliability of these constructs.

Accuracy. Accuracy was measured not only through functional features as declared by Ellis (2005), but also through semantic transparency. That is because an utterance can be error free and grammatically accepted but may lack communicative adequacy (Pallotti, 2009). In addition, phonological and morphological aspects including word choice were considered following Wulff & Gries’ (2011) suggestions. Phonological aspects were measured through punctuation, intonation, and stress. Morphological notions were measured by the correct use of closed class functional words such as articles, conjunctions, prepositions and pronouns, and the correct use of derivational and inflectional affixes of bond morphemes. The AS-unit of analysis was used to calculate the number of error free clauses in comparison to the number of clauses uttered. (Ellis, 2005; Skehan, 2009).

Fluency. When assessing fluency as producing language without pausing or interruption, it is vital to identify which of the different sub-types of fluency (breakdown fluency, repair fluency, and speed fluency) is being measured (Skehan, 2005; Pallotti, 2009). In this scheme, fluency was measured in terms of repair fluency which involves false starts and repetition (Pallotti, 2009). Since this scheme was developed on a communicative approach, fluency was measured through the AS-unit communicatively by counting the number of pauses that occurred as a result of lack of felicity with the total of speaking time (Skehan, 2009).

Complexity. Complexity has been identified and measured in diverse ways due to its complex nature in involving many aspects such as difficulty, development and other production factors of a learner's performance (Pallotti, 2009). The current scheme considered the comment addressed by Pallotti (2009) which suggested segregating development from complexity, while considering the product of the performance instead of the process. Following Ellis (2005), complexity in this scheme was measured by calculating the ratio of clauses compared to the subordinate clauses used (following the AS-unit of assessment). Unlike Ellis, lexical variety was not measured under complexity, but it was dealt with as a separate construct, as done in the study of Skehan (2009).

Lexis. According to Skehan (2009), Lexis should be considered as a separate construct and not under complexity or accuracy due to its importance in reflecting background knowledge of the L2 and its valid effect on performance. Therefore, Lexical adequacy and lexical variety was considered in this scheme as a separate construct. Lexical adequacy was measured communicatively in terms of felicity (Pallotti, 2009), and lexical variety was measured through anaphoric, cataphoric and exophoric elements (Skehan, 2009). The number of error free word choices and word variety was calculated in comparison to the number of words uttered following the AS-unit of analysis (Foster et al., 2000).

Communicative Adequacy. Defining and measuring meaning requires differentiating between explicit and implicit meanings (Purpura, 2017). All the previous main constructs in this scheme were measured based on communicative aspects implicitly. Communicative adequacy was also considered explicitly as a separate construct due to its importance in assessing language performance (Pallotti, 2009). Assessing communicative adequacy in this scheme explicitly was measured through coherence, which Coherence refers to the semantic transparency of the utterance or the implied rhetorical meaning (Purpura, 2017). Explicit communicative adequacy was assessed using the idea-unit as done by Ellis and Barkhuizen (2005) in their study. However, cohesive devices were measured through accuracy as mentioned above.

Scoring Criteria. The scoring Criteria was based on error analysis. Every construct was measured according to the AS unit or the idea-unit as mentioned previously. The ratios of error free words, clauses or pauses were calculated. The ratio out of a 100% was assigned for each construct. The mean of the total score was calculated as explained in the analysis of this study.

The TOEFL rubric. The TOEFL speaking rubric (ETS, 2014) was used in this study to test the compatibility of the current scheme in comparison. According to Jamieson and Poonpon (2013), the TOEFL speaking rubric has been subjected to many tests and is considered to be a valid and reliable assessment tool, based on the communicative approach to language learning. The general categories of the rubric involved three main constructs which were: delivery, language use, and topic development. The scores were based on a holistic scoring system. The following will explain components of both delivery and language use and the scoring system which was applied.

Delivery. The delivery category involved aspects of fluency. It measured fluency through clarity, lapses, pronunciation, intonation, intelligibility, pacing and flow of speech. The highest proficiency level (level 4) involved well-paced flow, clear speech production, minor lapses, minor difficulties in pronunciation, and minor difficulties in intonation patterns. However, all of these minor mistakes should not affect the comprehensibility of the speech production. The level of proficiency lessened according to the performance of the participants in achieving these notions (as shown in Appendix 2).

Language Use. Aspects of accuracy, complexity, and lexis were considered here. Accuracy was measured through effective use of grammar rules and lexical adequacy. Complexity was measured by using basic and complex structures. Accuracy, Complexity, and lexis were measured through communicative purposes. The highest level of proficiency in language use may contain minor mistakes as long as it does not affect the meaning conveyed (see Appendix 2). The overall idea is considerable depending upon the levels of proficiency.

Topic Development. Topic development was measured through sustainability of ideas, and sufficiency of task requirements. Coherence was measured through clear progression of ideas. The least proficiency level involved aspects of irrelevance of ideas and lack of elaboration (see Appendix 2 for details).

Communicative adequacy. TOEFL speaking rubric was mainly based on the communicative approach of language learning (Jamieson & Poonpon, 2013). All three categories in the TOEFL rubric were assessed according to the meaning of language that the participant conveyed. Communicative adequacy was measured implicitly in all of the categories, rather than measuring it explicitly as a separate construct.

Scoring Criteria. These categories were measured to determine the level of proficiency from zero to four. The highest performance level in all three categories (delivery, language use, and topic development) was allocated with a score of four and the least performance is allocated with a score of one. If the participant is off-topic or did not respond, a score of zero was assigned. The means of the total score were calculated as explained in the proposed analysis of this study.

The Instruments Reliability and validity. The task-based activity adapted from Ellis (2012) and the TOEFL rubric have been subjected to various reliability and validity measures (Ellis, 2012; Jamieson & Poonpon, 2013). Therefore, they were not tested in this study. However, the CAF scheme was subjected to reliability measures by investigating repeatability and reproducibility (Allen & Knight, 2009). To ensure repeatability, the researcher piloted the scheme by assessing another sample (n=5) and after a week, the same sample was reassessed using the same rubric to ensure there were no differences in students' scores when assessed again. As for reproducibility, another examiner holding the same proficiency level of the researcher, assessed the same students using the same rubric to ensure no differences were noted.

To validate the scheme several aspects of validity were considered following Mackey and Gass (2016) such as, face validity, content validity, and construct validity. Face validity was tested by the general appeal of the scheme which appears to test what it is meant to be tested. Content validity was investigated through previous research suggestions that the scheme was based on. Then, construct validity was achieved through this study when the CAF scheme was compared to the validated TOEFL rubric following Jamieson and Poonpon (2013).

2.5 Study Participants

This study involved 30 female university students in the College of Languages and Translation at King Saud University in Riyadh Saudi Arabia. The selected participants belonged to level three of the language speaking course in English as a foreign language. The average age of the selected participants was 20 years. The rationale of choosing KSU students is the lack of a unified assessment tool which measures their performance in speaking communicatively.

3.0 PROCEDURE

The procedure of the data collection followed three main stages (see Table 2). The participants were given a task-based activity adapted from Ellis (2012). The instructions were stated clearly by the researcher. Then the students were given time to read and understand the task. After that, the participants were required to work in pairs to answer the task. The responses of the participants were recorded in a lab while they were answering the task. Each participant had a lapel to record their responses clearly and individually following Mackey and Gass (2016). Then the same researcher assessed the participants' performance for the second time using different assessment schemes. First the responses of the participants were evaluated according to the CAF scheme. The mean of the total score was calculated. After that, the researcher evaluated the performance of the same participants using the TOEFL rubric, while calculating the mean of the total score. After that the means of both tests were statistically tested as explained in the following section.

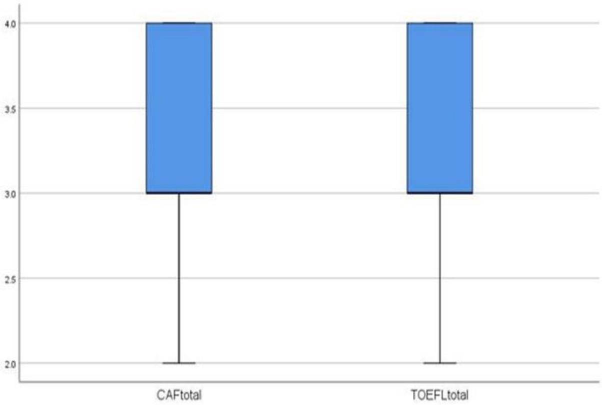
Table 2. The Sequence of the Data Collection of the Study

Stages	Procedure
Stage 1	The participants answered a task-based activity and their responses were audio recorded in a lab with individual lapels).
Stage 2	The performances of the participants were assessed using the speaking assessment scheme based on reconstructed CAF measures. The mean of the total score was calculated.
Stage 3	The performances of the participants were assessed using the TOEFL rubric. The mean of the total score was calculated.

3.1 Data Analysis

The data collected was analyzed quantitatively to answer the research question of this study. Prior to choosing the suitable statistical test, the data collected was tested for normality. The purpose was to test if the data collected was normally distributed in order to choose the suitable parametric or nonparametric statistical test. Using the SPSS, the normality test was shown through a descriptive statistical table and QQ plots. The statistical tests of Kolmorov-Smirnov, and Shapiro-Wilk were used. According to which if the p-value is less than 0.5, the null hypothesis is attained which means that the data normally distributed is rejected (Larson hall, 2016). According to table 3, the p-value is less than 0.5, therefore, revealing that the data is not normally distributed. However, the QQ plots and the box plot shows otherwise (figure 1).

Figure 1. Plots of Normality



According to Larson hall, a normal distribution may not be clear from the tests alone and that if the QQ plots show normal distribution, a parametric statistical test may be used (2016). Therefore, this study will follow Larson hall and conclude that the data is normally distributed and a parametric statistical test was used.

Table 3. Test of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
CAFtotal	.285	30	.000	.789	30	.000
TOEFLtotal	.292	30	.000	.772	30	.000

a. Lilliefors Significance Correction

Furthermore, a parametric statistical paired t-test was used to analyze the mean score of two assessments of the same group (Larson Hall, 2016). The paired t-test was conducted by using SPSS. The p-value of the t-test was calculated to determine if there were statistically significant differences between the two assessments. The means, effect size, and the 95% confidence intervals were reported following Larson-Hall (2016) to show the size of the results concluded and their significance. The power was calculated post-hoc using the statistical program R.

4.0 FINDINGS AND DISCUSSION

The performances of the participants (N= 30) were assessed twice. First, they were assessed by using the Speaking assessment schemes based on suggestions from the literature to reconstruct CAF measures. Then, they were assessed by using the TOEFL speaking Rubric. The mean scores of the participants performance on both tests (M= 3.2; M=3.3) and the standard deviation (Std. D= .66; Std. D = .71) were similar (as shown in Table 4).

Table 4. Mean Scores of participants on CAF and TOEFL speaking Rubrics

	Mean	*N*	*Std. Deviation*	*Std. Error Mean*
RCAF	3.2000	30	.66436	.12130
TOEFL	3.3333	30	.71116	.12984

Moreover, the strength of the correlation between the results of the two tests should be large or otherwise, it would be problematic due to the necessity of having a correlation in a paired sample t-test (Larsonhall, 2017). Hence, the strength of the correlation has been calculated by the parametric paired t-test, as shown in table 5. The result indicated that there was a strong correlation between both variables (r = .876) at a significance as low as 0.0.

Table 5. Paired Samples Correlation

	N	*Correlation*	*Sig.*
RCAF	30	.876	.000
TOEFL			

Furthermore, according to the parametric paired t- test, the t-score calculated as a statistic for this test has a probability below p = .05 which means that the probability of finding a paired t-test is strong, if the null hypothesis were true i.e. p = 0.43. Therefore, the results of this study supported this hypothesis. In table 6, the first column indicates the calculated mean of the two tests (M =-.13) followed by the standard deviation (std. D = .34575). The 95% Confidence Intervals for the mean difference between the two tests (CI = -.26; CI = -.00) were achieved between -.26 and -.00. This means that the difference between the mean scores could be as large as .26 or as small as zero with 95% confidence. This difference is very slight and it is quite close to zero. Findings indicate that there is no difference between the results of the participants on both tests.

Table 6. Paired sample t-test

	Paired Differences							
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
				Lower	Upper			
RCAF	-	.34575	.06312	-.26244	-.00423	-2.112	29	.043
TOEFL	.13333							

In addition, the effect size was calculated through Cohen's d by using the t-value and the degrees of freedom in the following equation d = 2t/df. The effect size of this test is -0.8 which is a very small effect according to Cohen's guidelines (Larsonhall, 2016). Power was calculated post hoc using the statistical program R. The power calculated was 0.07 which is considered to be very low. Therefore, findings of the study indicated that hypothesis which stated no statistically significant differences in the means of the total scores of the participants on the CAF scheme compared to their results on the TOEFl rubric is true and is supported in this study.

This study tested a speaking assessment rubric based on literature suggestions to reconstruct CAF measures (Skehan, 2009; Palloti, 2009). It tested its compatibility to the TOEFL speaking rubric (ETS 2014). The CAF scheme followed the literature suggestions in measuring performance in Speaking through five independent constructs that includes Lexis (Skehan, 2009), Fluency, Accuracy, Complexity (Ellis, 2012; Wulff & Gries, 2011), and Communicative adequacy (Pallotti, 2009; Purpura, 2017).

Findings of this study indicated that there were no statistical differences between the results on both tests. Since the TOEFL rubric is a valid and reliable assessment tool, comparing the results of the CAF speaking assessment to the results of TOEFL, it has ensured that the RCAF rubric has construct validity. Furthermore, the small power that has been calculated post hoc through the statistical program R may be the reason behind not being able to indicate a normal distribution in the normality tests used through Kolmogorov-Smirnov and Shapiro-Wilk (Larsonhall, 2016).

5.0 CONCLUSION AND IMPLICATIONS

This study intended to synthesize research on the concept of reconstructing CAF and its significance in assessing second language performance through shedding light on how CAF was identified and operationalized in research, and the implications mentioned to minimize the limitations of CAF. Also, this study attempted to test a speaking assessment scheme based on previous research suggestions to reconstruct CAF measures by comparing it to the TOEFL speaking rubric (ETS, 2014). Score means were analyzed and discussed. Study findings supported the hypothesis which stated that there were no statistical differences in the results of the participants in both assessment schemes.

This study focused on the mean of the total score of both tests. Future researches may investigate the parts of both rubrics and if there are significant differences between them. Also, a major limitation in the scheme was the measurement method which was calculating the number of words uttered and the number of main and subordinate clauses. It was time consuming and difficult to count at times. Therefore, future research can use an automated application that can transcribe and compute these measures instead of calculating them manually.

5.1 Acknowledgment

The author is indebted to the two anonymous reviewers for their insightful and helpful comments. The authors also wish to express their gratitude to both the Deanship of Scientific Research and the Research Centre at the Faculty of Language Sciences for funding the current article. Also, the authors thank RSSU at King Saud University for their technical support.

REFERENCES

Allen, S., & Knight, J. (2009). A Method for Collaboratively Developing and Validating a Rubric. International Journal for the Scholarship of Teaching and Learning, 3(2), n2. https://doi.org/10.20429/ijsotl.2009.030210

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2e éd.). Newbury Park, É. U. Sage. https://doi.org/ 10.1016/b978-0-12-179060-8.50006-2

De Jong, N., & Vercellotti, 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. https://doi.org/10.1177/1362168815606161

Ellis, R. (2012). Language teaching research and language pedagogy. John Wiley & Sons. https://doi.org/ 10.1002/9781118271643

Ellis, R., & Barkhuizen, G. P. (2005). Analysing learner language. Oxford: Oxford University Press.

ETS. (2014). TOEFL iBT test-Independent Speaking Rubric (Scoring Standards). Priceton, NJ: Author. Retrieved from https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. Studies in Second language acquisition, 18(3), 299-323. https://doi.org/ 10.1017/s0272263100015047

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. Applied linguistics, 21(3), 354-375. https://doi.org/10.1017/s0272263100015047

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. Applied linguistics, 30(4), 461-473. https://doi.org/10.1093/applin/amp048

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. https://doi.org/10.1075/lllt.32.01hou

Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3.

Jamieson, J., & Poonpon, K. (2013). Developing Analytic Rating Guides for Toefl Ibt's Integrated Speaking Tasks. ETS Research Report Series, 2013(1), i-93. https://doi.org/ 10.1002/j.2333-8504.2013.tb02320.x

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. Applied linguistics, 27(4), 590-619. https://doi.org/10.1093/applin/aml029

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. Applied linguistics, 30(4), 579-589. https://doi.org/10.1093/applin/amp043

Larson-Hall, J. (2015). A guide to doing statistics in second language research using SPSS and R. Routledge. https://doi.org/ 10.4324/9780203875964

Mackey, A., & Gass, S. M. (2016). Second language research: methodology and design. New York and London. https://doi.org/ 10.4324/9781410612564

Norris, J. M., & Ortega, L. (Eds.). (2006). Synthesizing research on language learning and teaching (Vol. 13). John Benjamins Publishing. https://doi.org/ 10.1075/lllt.13

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. Applied linguistics, 30(4), 590-601. https://doi.org/10.1093/applin/amp045

Purpura, J.E. (2017). Assessing Meaning: Language Testing and Assessment. In Encycolpedia of Language and Education, pp. 33-61. https://doi.org/10.1007/978-3-319-02261-1_1

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. Applied linguistics, 22(1), 27-57. https://doi.org/10.1093/applin/22.1.27

Sample, E., & Michel, M. (2014). An exploratory study into trade-off effects of complexity, accuracy, and fluency on young learners' oral task repetition. TESL Canada Journal, 23-23. https://doi.org/ 10.18806/tesl.v31i0.1185

Skehan, P. (1998). A cognitive approach to language learning. Oxford University Press. https://doi.org/10.1177/003368829802900209

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. Applied linguistics, 30(4), 510-532. https://doi.org/10.1093/applin/amp047

Vercellotti, M. L. (2012). Complexity, accuracy, and fluency as properties of language performance: The development of multiple subsystems over time and in relation to each other (Doctoral dissertation, University of Pittsburgh).

Wulff, S., & Gries, S. T. (2011). Corpus-driven methods for assessing accuracy in learner production. Second language task complexity: Researching the cognition hypothesis of language learning and performance, 61, 87. https://doi.org/10.1075/tblt.2.07ch3