Article

Not peer-reviewed version

Syntactic Complexity and Proficiency: A Study of Objective Measures in the TEEP Speaking Test

Munis Akilova *

Posted Date: 23 January 2025

doi: 10.20944/preprints202501.1701.v1

Keywords: TEEP speaking test; proficiency; syntactic complexity; validity; objective measures; assessment

Article

Syntactic Complexity and Proficiency: A Study of Objective Measures in the TEEP Speaking Test

Munis Akilova

Department of Foreign Languages, Tashkent State University of Law (TSUL), Shakhrisabz Street, 33, Tashkent 100000, Uzbekistan; akilova@gmail.com; Tel.: +998 71 123 4567

Abstract: Syntactic complexity plays a critical role in assessing language proficiency, yet identifying measures that effectively distinguish proficiency levels remains challenging. This study investigates the relationship between objective measures of syntactic complexity and subjective ratings in the Test of English for Educational Purposes speaking assessment. Focusing on three metrics Mean Length of AS-Unit, Mean Length of Clause, and Subordinate Clauses per AS-Unit the research evaluates their ability to correlate with subjective ratings of grammatical range and accuracy and differentiate proficiency levels. Using a quantitative approach, the study analyzed 89 TEEP speaking test transcriptions paired with proficiency scores ranging from 5.0 to 7.5. Statistical analyses revealed weak but significant positive correlations between MLAS and MLC with subjective ratings, suggesting that these measures contribute modestly to evaluative outcomes. However, SCP-AS showed no significant correlation. ANOVA results highlighted that while MLAS distinguished between lower and higher proficiency levels, it struggled to differentiate adjacent levels. MLC and SCP-AS showed limited variation across levels, indicating low sensitivity. The findings suggest refining TEEP rating scales to assess complexity and accuracy separately and incorporating diverse syntactic, lexical, and morphological measures to enhance assessment validity. Future research should expand metrics and task types to deepen understanding of syntactic complexity in language assessment contexts.

Keywords: TEEP speaking test; proficiency; syntactic complexity; validity; objective measures; assessment

Introduction

The dimensions of complexity, accuracy, and fluency (CAF) are fundamental in second language acquisition (SLA) research for analyzing learners' written and oral performance as well as their overall proficiency. These dimensions are characterized as "indicators of learners' proficiency underlying their performance" (Housen & Kuiken, 2009, p. 461). Since the 1990s, complexity has been acknowledged as a core component of language proficiency, reflecting a learner's knowledge of a second language (L2). It holds a pivotal role in the speaking construct described by Fulcher (2009). Prominent international English proficiency exams, including the TOEFL iBT, IELTS, and OET, have incorporated complexity into their assessment frameworks, either as a standalone criterion or as part of broader categories such as grammatical range and accuracy or lexical resource.

Evaluating complexity in language performance encompasses various dimensions, including interactional, propositional, grammatical, lexical, and functional aspects (Ellis & Barkhuizen, 2005). Among these, syntactic complexity, which refers to the intricacy of sentence structures and grammatical constructions, is particularly significant for assessing language proficiency. A range of objective measures has been introduced to evaluate syntactic complexity, with higher proficiency levels generally associated with greater complexity in language use. However, the relationship between specific measures of complexity and proficiency levels may vary depending on test tasks, the language skills being assessed, and proficiency descriptors (Ellis & Barkhuizen, 2005; Skehan &

Foster, 1997). Proficient speakers are expected to demonstrate effective use of more complex sentence structures (Kuiken, 2021; Kuiken & Vedder, 2012). Despite this, identifying measures that consistently capture complexity across different proficiency levels remains a challenge.

The Test of English for Educational Purposes (TEEP), developed by Cyril Weir at the University of Reading in the 1980s, builds upon this understanding of complexity in language assessment. The TEEP is designed to assess English language proficiency for academic purposes, determining whether students possess the necessary language skills to undertake higher education in the UK. Over the years, the TEEP has been administered not only in the UK but also in countries such as Malaysia and China, where it has evaluated the language abilities of thousands of undergraduate and postgraduate students. The TEEP's language assessment employs both holistic and analytic ratings, with raters trained to evaluate syntactic complexity. While existing research highlights the importance of combining objective measures, such as lexical and syntactic analyses, with subjective ratings for assessing language complexity, the relationship between these measures and their effectiveness in distinguishing proficiency levels within the TEEP remains insufficiently explored.

This study seeks to investigate how syntactic complexity is conceptualized and assessed within the TEEP test and whether subjective evaluations accurately reflect the linguistic complexity of test takers' speech across varying proficiency levels. By analyzing the relationship between objective and subjective assessments of syntactic complexity, this research aims to enhance the reliability and validity of spoken English proficiency evaluations within the TEEP framework. Insights gained from this study may contribute to more effective assessment practices and provide better support for learners' language development.

2. Literature review

2.1. Conceptualizing L2 Complexity in SLA

Complexity is widely regarded as one of the most challenging constructs within the CAF (Complexity, Accuracy, Fluency) framework due to its multidimensional nature (Bulté & Housen, 2012; Housen & Kuiken, 2009; Pallotti, 2009). In SLA research, "complexity" is used to describe both task-related characteristics (task complexity) and aspects of second language performance and proficiency (L2 complexity). To address this ambiguity, researchers have proposed using "difficulty" to describe task-related characteristics, reserving "complexity" to indicate language development as a dependent variable rather than an independent factor that influences task difficulty or cognitive demands (Kuiken & Vedder, 2012; Pallotti, 2009; Skehan, 2006). However, even when applied solely to performance, complexity retains multiple meanings, as it encompasses various linguistic and communicative dimensions (Ellis & Barkhuizen, 2005; Housen & Kuiken, 2009; Housen, Kuiken, & Vedder, 2012).

Ellis and Barkhuizen (2005) outline eight dimensions of complexity, including propositional, interactional, lexical, and grammatical complexity. Similarly, Bulté and Housen (2012) argue that L2 complexity encompasses three components: propositional complexity, discourse-interactional complexity, and linguistic complexity. Propositional complexity involves the quantity of information units encoded in a message, while discourse-interactional complexity is concerned with features such as turn-taking, interactional moves, and participation roles in dialogic discourse. Despite their relevance, these aspects have received considerably less attention than linguistic complexity (Kuiken, 2021).

Linguistic complexity can be understood at two levels: as a dynamic property of the learner's overall interlanguage system and as a stable property of specific linguistic features within that system (Bulté & Housen, 2012; Housen & Kuiken, 2009; Vercellotti, 2015). At the interlanguage level, it refers to the size, elaborateness, richness, and diversity of the learner's L2 system. At the feature level, structural complexity is further divided into formal and functional aspects. Linguistic complexity is often categorized into grammatical complexity, which pertains to syntactic and morphological structures, and lexical complexity, which encompasses lexical density (the proportion of lexical words), sophistication (the use of infrequent words), and diversity (the variety of words used) (Bulté & Housen, 2012).

2.2. Defining L2 Complexity

The multifaceted nature of complexity has led to the absence of a universally accepted definition in SLA research. This lack of clarity has resulted in varying interpretations and inconsistent findings in complexity studies (Bulté & Housen, 2012). Wolfe-Quintero et al. (1998) define grammatical and lexical complexity as a learner's ability to use a broad range of basic and advanced structures and vocabulary. Ellis and Barkhuizen (2005) describe it as the capacity to produce more elaborate and challenging language. Skehan (1998) links complexity to advanced language use and a learner's willingness to take risks with unfamiliar language.

Pallotti (2009) highlights linguistic variation as a key dimension of complexity but questions the validity of using developmental progress as a complexity indicator. Some linguistic forms, such as the subjunctive mood or inversion in conditionals, may be acquired later not due to their structural complexity but because of their low frequency or limited communicative relevance. Thus, Pallotti argues for distinguishing between complexity as linguistic production and development as part of the learning process. She categorizes complexity into three types: structural complexity, which pertains to the number of elements and relationships within linguistic systems; cognitive complexity, which relates to the processing demands of linguistic structures; and developmental complexity, which refers to the order in which linguistic forms are acquired (Pallotti, 2015).

This study adopts a structural perspective on complexity, focusing specifically on syntactic complexity. Following Bulté and Housen's (2012) definition, complexity is understood as the number of distinct components in a language feature and the systematic relationships among them. In this context, L2 complexity is typically analyzed by examining the quantity and variety of lexical items and syntactic structures, the length of these structures, and the degree of embedding or dependency relationships. A detailed discussion of these measures will follow in subsequent sections.

2.3. Measures of Syntactic Complexity

In SLA research, syntactic complexity has been operationalized through various measures, including the length of utterances, diversity of syntactic structures, and sophistication of these structures. These measures are commonly categorized into general (or global) and specific metrics. General measures often emphasize length and subordination. Length-based metrics calculate the ratio of word frequency to a specific syntactic unit (Bulté & Housen, 2012; Kuiken, 2023).

One of the most widely used units in SLA studies is the minimal terminable unit (T-unit), which is defined as a main clause along with its dependent clauses and any attached non-clausal units (Norris & Ortega, 2009; Hunt, 1965, as cited in Ehret et al., 2023). However, applying T-units to spoken language presents challenges, particularly in segmenting data, as they were originally designed for analyzing L1 written production. The presence of dysfluencies and interruptions in spontaneous speech often complicates the use of T-units, leading to concerns about their applicability (Foster et al., 2000; Ellis & Barkhuizen, 2005). To address these issues, alternative units such as c-units, defined as "utterances, words, phrases, grammatical and ungrammatical, that provide pragmatic or referential meaning" (Bardovi-Harlig, 1992, p. 54), and AS-units, defined as "a single speaker's utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clauses associated with it" (Foster et al., 2000, p. 4), have been proposed.

Another group of syntactic complexity metrics focuses on subordination, quantifying the number of clauses in relation to a production unit, such as the mean number of clauses per T-unit, c-unit, or AS-unit. Bardovi-Harlig (1992) suggested the coordination index, which measures the proportion of coordinated clauses relative to the total number of clauses, as a more sensitive indicator of complexity during early stages of L2 development.

Specific metrics, on the other hand, assess advanced grammatical structures, including the frequency of infinitive phrases, auxiliary verbs, passive constructions, imperatives, comparatives, and conditionals. These features are often associated with more advanced learners (Ellis & Barkhuizen, 2005; Norris & Ortega, 2009).

2.4. Complexity in Language Proficiency Assessments

Language proficiency levels in assessment frameworks are detailed through sub-constructs and rubrics, where complexity is incorporated into a global proficiency rating and evaluated alongside other dimensions such as accuracy, fluency, pronunciation, and communicative effectiveness. Table 1 provides an overview of linguistic features used to assess speaking in TEEP, IELTS, and TOEFL iBT.

Table 1. Assessment of Speaking in TEEP, IELTS, TOEFL iBT.

TEEP	IELTS	TOEFL iBT
Fluency and coherence	Fluency and coherence	Delivery
Lexical resource	Lexical resource	Language use
Grammatical range and accuracy	Grammatical range and accuracy	Topic development
Pronunciation	Pronunciation	-
Interactive communication	-	-

As illustrated in Table 1, TEEP, IELTS, and TOEFL iBT consider complexity within their rubrics, addressing lexical, syntactic, and discourse complexity within broader categories. Lexical complexity is assessed by analyzing the range and appropriateness of vocabulary, while discourse complexity is evaluated based on the organization of ideas, coherence, and use of cohesive devices. Syntactic complexity, on the other hand, reflects the ability to utilize varied and advanced grammatical structures, with higher proficiency levels typically demonstrating greater complexity.

Despite the inclusion of lexical, syntactic, and discourse complexity in their frameworks, these tests (TEEP, IELTS, and TOEFL iBT) often fail to differentiate clearly between syntactic complexity and grammatical accuracy, merging the two into a single assessment criterion. This conflation can present challenges for both test-takers and raters, as candidates may exhibit strength in one aspect while struggling with the other, making it difficult to distinguish between varying levels of performance. Additionally, rating scales may lack the precision needed to separate different degrees of syntactic complexity and accuracy, which could undermine scoring consistency and affect the reliability and validity of the assessments. Addressing this issue is essential to ensure that proficiency evaluations provide an accurate and comprehensive representation of the multifaceted nature of language ability.

2.5. Research in L2 Speech Complexity and Proficiency

Numerous studies have examined the relationship between L2 speech complexity and proficiency. While some research has focused on languages like Japanese (Iwashita, 2006) and French (De Clercq, 2015; De Clercq & Housen, 2017), this study specifically investigates L2 English.

Brown et al. (2008) analyzed measures of accuracy, fluency, and complexity across five levels (A1 to C1) of the TOEFL iBT speaking test. Their study primarily focused on syntactic complexity, including the T-unit complexity ratio (clauses per T-unit), the dependent clause ratio, the verb-phrase ratio (verb phrases per T-unit), and the mean length of utterances. Significant differences between proficiency levels were observed in syntactic complexity measures, particularly in verb phrases per T-unit and mean utterance length.

Seedhouse et al. (2014) explored the relationship between performance in the IELTS speaking test and candidates' scores. They measured syntactic complexity using clause-linking metrics, such as subordinate clauses per AS-unit and per total number of words. Results showed that grammatical complexity did not consistently increase across band levels. Interestingly, candidates at Band 7 demonstrated greater grammatical complexity than those at Band 8.

Bulté and Roothooft (2020) investigated the relationship between nine quantitative measures of L2 speech complexity and IELTS proficiency levels, analyzing oral productions across five levels (4–8). Their analysis included measures like mean length of AS-units (MLAS), mean length of clauses (MLC), noun phrase length, subclause ratio, and coordination ratio. Four of these measures—MLAS, MLC, noun phrase length, and subclause ratio showed significant differences across proficiency levels, effectively distinguishing between adjacent levels, particularly Levels 6 and 7.

The studies collectively highlight that syntactic complexity measures, particularly those assessing the average length of supra-clausal units and clausal subordination, tend to increase with higher proficiency levels. However, variations in research methodologies, learner populations, and statistical approaches complicate generalizations. Additionally, a key limitation in previous research is the narrow focus on monologic tasks to assess speaking proficiency, despite the predominance of dialogic communication in natural language use. While some studies have explored the effectiveness of specific complexity measures in distinguishing proficiency levels, there remains a gap in comprehensively examining syntactic complexity across proficiency levels in diverse testing contexts, such as the TEEP speaking assessment. This study addresses these gaps by investigating the relationship between subjective ratings and objective syntactic complexity measures in dialogic tasks across TEEP proficiency levels. Specifically, it seeks to answer the following research questions:

1. To what extent do objective measures of syntactic complexity in dialogic tasks correlate with subjective ratings of range and accuracy in the TEEP speaking assessment?
2. To what extent can objective measures of syntactic complexity reliably distinguish between different proficiency levels in the TEEP speaking paper?

3. Methodology

3.1. Research Design

The study adopted a quantitative research design, a common approach in language testing and assessment research (Paltridge & Phakiti, 2015). This method is particularly suitable due to its efficiency, cost-effectiveness, and ability to handle large datasets systematically. Quantitative research allows for the statistical analysis of data, making findings replicable and generalizable (Dörnyei, 2007). In this study, it facilitated a rigorous examination of test-takers' spoken language, enabling precise measurement of syntactic complexity and its correlation with proficiency levels in the TEEP speaking test.

By relying on numerical data and statistical analysis rather than subjective interpretation, the study minimized potential biases, thereby enhancing the reliability and validity of its conclusions (Dörnyei, 2007). This objectivity provided clear evidence of relationships between syntactic complexity and proficiency levels, ensuring robust results.

In recent years, a data-driven approach has been increasingly employed in language assessment research to analyze key features of test-taker performances, validate score descriptors, and distinguish between adjacent proficiency levels (Tavakoli et al., 2017). This method has been widely used to investigate spoken language features across different proficiency levels in high-stakes tests, such as IELTS (Seedhouse et al., 2014), TOEFL iBT (Brown et al., 2006), and Cambridge English (Kang, 2013). Similarly, the present study aimed to explore the relationship between TEEP subjective ratings and objective measures of syntactic complexity across various proficiency levels by analyzing the oral responses of test-takers. The following sections provide detailed explanations of the research process.

3.2. Data Set

This study was conducted in collaboration with the Assessment Team at the International Study and Language Institute of the University of Reading, which provided the necessary data. The dataset included transcriptions of TEEP Speaking Test responses and candidates' proficiency scores, sourced from an existing database. The researchers express their gratitude to the Assessment Team for their valuable contribution.

The TEEP Speaking Test is administered in pairs and consists of three interconnected sections designed to assess candidates' ability to communicate effectively in an academic context. The test topics cover a range of subjects, including crime, education, international aid, language and culture, tourism, and health. Table 2 outlines the structure of the TEEP Speaking Test.

Table 2. Structure of TEEP Speaking Test.

Part	Task	Mode	Description	Planning time	Response time
1	Focus/topic introduction	Silent preparation	Examinees prepare silently before responding to a prompt question	20 seconds	—
2	Individual talk (role plays)	Monologue	Examinees engage in role plays discussing advantages or disadvantages of a given topic	4 minutes	3 minutes
3a	Scenario discussion	Dialogue	Examinees discuss specific scenarios	2 minutes	4 minutes
3b	Further discussion	Dialogue	Examinees analyze and discuss the focus question	—	—

This study focused exclusively on Task 3, as it is longer than the other tasks, demands higher cognitive effort, and involves dialogic interaction. In Task 3a, candidates are presented with a visual prompt that outlines a problem and provides three potential solutions. After two minutes of preparation, they discuss the scenario, evaluate the options, and demonstrate interactional competence through active language use. For test security reasons, specific task contents were not disclosed.

The dataset included 89 transcriptions of Task 3a responses, paired with candidates' TEEP speaking proficiency scores. These scores, ranging from 5.0 to 7.5, were assigned by TEEP raters using a standardized rating scale (see Chapter 2 for scoring procedures). Each transcription was matched with its corresponding score, ensuring a consistent analysis of syntactic complexity in relation to proficiency. Table 3 summarizes the distribution of performances across proficiency levels.

Table 3. Distribution of Speech Performances Across Proficiency Levels.

PROFICIENCY LEVEL	NUMBER OF PERFORMANCES
5.0	13
5.5	12
6.0	9
6.5	15
7.0	19
7.5	21
TOTAL	89

3.3. Participants

The dataset was drawn from a diverse cohort of English language learners who undertook the Test of English for Educational Purposes (TEEP) across various proficiency levels. The participants, comprising both male and female test-takers, represented a range of first language (L1) backgrounds and nationalities. Their proficiency levels, reflected in overall TEEP scores, ranged from 5.0 to 7.5, including half-band scores. These levels were chosen to encompass a broad spectrum of English proficiency, from elementary to advanced, enabling a comprehensive analysis of syntactic complexity across skill levels. The inclusion of half-band scores was particularly relevant, as these are significant for institutional progression requirements and often present challenges for raters due to their nuanced nature in the TEEP scoring scale.

3.4. Data Preparation and Coding

The transcriptions included detailed features of spoken language, such as hesitations, repetitions, self-corrections, and false starts. During data preparation, hesitation markers (e.g., "uh," "erm") and minor utterances (e.g., "yeah," "okay," "well") were removed. False starts (e.g., "I watch, I played chess yesterday") were revised to their final form ("I played chess yesterday"), and self-corrections and repetitions (e.g., "She like to, likes to read books") were standardized (e.g., "She likes to read books").

The samples were divided into AS-units and clauses, as defined in Chapter 2. Following Foster et al. (2000), subordinate clauses were treated as separate clauses if they contained a finite or non-finite verb and at least one additional element (e.g., subject, object, or adverbial). For instance, "It is a good place to visit" was treated as a single clause, while "They want to live in Italy" was divided into two clauses. Coordinated main clauses were analyzed as separate units, while coordinated verb phrases with the same subject were considered part of a single unit. Interruptions were addressed by noting the precise point where a speaker's utterance was cut off. Interrupted segments were classified as incomplete AS-units, with the analysis focusing on the uninterrupted portions of the discourse (Foster et al., 2000).

The coding process involved two coders: the researcher and a Master's student in TESOL with relevant expertise. To ensure consistency and accuracy in applying the coding criteria, three workshops were conducted under the guidance of an expert in the field. During these workshops, 29 transcriptions were independently coded by both coders and reviewed to resolve discrepancies and establish standardized procedures. Afterward, the remaining transcriptions were split, with each coder independently coding 30 samples. Additionally, three double-coding meetings were held to discuss and reconcile any remaining inconsistencies. Although formal statistical measures of inter-coder reliability were not utilized, the collaborative and supervised process aimed to ensure consistent and reliable coding outcomes.

3.5. Complexity Measures

The study employed three syntactic complexity measures, as summarized in Table 4, to analyze the data.

Table 4. Complexity measures.

Complexity measures	**Calculation**
Mean length of AS unit (MLAS)	Total number of words divided by total number of AS units
Mean length of clause (MLC)	Total number of words divided by total number of clauses
Subordinate clauses per AS-unit	Total number of clauses divided by total number of AS units

Note: These measures are adapted from Norris and Ortega (2009).

These measures were selected for several reasons. Firstly, they have a longstanding tradition in applied linguistics and have been validated in numerous studies (Bardovi-Harlig, 1992; Brown et al., 2005; Ortega, 2003; Iwashita et al., 2006; Seedhouse, 2017). Their consistent use across studies ensures comparability and reliability.

Secondly, they provide valuable insights into syntactic complexity. MLAS and MLC capture sentence length and complexity, indicating learners' ability to construct longer, more intricate sentences (Ellis & Barkhuizen, 2006). Subordinate Clauses per AS-unit, on the other hand, highlights the use of subordination, a critical indicator of syntactic development in intermediate and advanced learners (Norris & Ortega, 2009; Skehan, 2009).

Lastly, these measures are particularly relevant to the TEEP Speaking Test. Given the test's aim of assessing a range of proficiency levels, these metrics effectively differentiate between levels by revealing syntactic features associated with different stages of language acquisition. Although more detailed measures could provide additional insights, the selected metrics balance practicality with comprehensiveness, making them suitable for the study's scope.

3.6. Data Analysis Tools

The analysis of the coded data was conducted using SPSS software, focusing on the relationship between syntactic complexity and proficiency levels in the TEEP Speaking Test. Descriptive statistics were used to calculate mean values for each complexity measure, and several inferential statistical techniques were employed to address the research questions.

For **RQ1**, which investigated the correlation between syntactic complexity measures and subjective ratings of range and accuracy, Pearson correlation analysis was conducted. Syntactic complexity measures (MLAS, MLC, Subordinate Clauses per AS-unit) and subjective ratings were entered into SPSS, and Pearson's correlation coefficient was used to evaluate the strength and significance of the relationships. The resulting coefficients and p-values provided insights into how closely the syntactic complexity measures aligned with subjective ratings.

For **RQ2**, which examined whether syntactic complexity measures could reliably differentiate between proficiency levels, a one-way ANOVA was performed in SPSS. Separate ANOVA tests were conducted for each complexity measure to identify significant differences across proficiency levels (5.0, 5.5, 6.0, 6.5, 7.0, 7.5). Since ANOVA does not indicate which levels differ significantly, Bonferroni post-hoc comparisons were conducted to pinpoint significant differences between proficiency groups.

4. Analysis and Results

4.1. Introduction

This chapter presents the findings from the data analysis conducted in this study, which utilized SPSS software as outlined in Chapter 3. The analysis begins with a descriptive overview of the dataset, followed by the results of inferential analyses (Pearson correlation and ANOVA) to address the research questions. Each research question is examined in turn.

Prior to conducting the inferential analyses (correlations and ANOVAs), descriptive statistics were calculated to summarize key features of the dataset. This involved determining the mean, standard deviation, range, and sample size (N) for each variable. Table 5 below provides an overview of these statistics, offering insight into the primary variables explored in the study.

Table 5. Descriptive Statistics.

N		Minimum	Maximum	Mean	Std. Deviation
MLAS	89	6.87	19.20	12.2803	2.34848
MLC	89	4.23	10.00	6.5671	1.07250
SCP-AS	89	1.17	3.42	1.8920	.40881

Proficiency Level	89	5.0	7.5	64.38	8.849

Note: MLAS = Mean Length of AS-unit; MLC= Mean Length of Clause; SCP-AS = Subordinate Clauses Per AS-unit.

The MLAS ranged from 6.87 to 19.20, with an average value of 12.28 words per AS-unit and a standard deviation of 2.35, indicating moderate variability in sentence length among participants. Similarly, MLC values varied between 4.23 and 10.00, with a mean of 6.57 and a standard deviation of 1.07, reflecting moderate consistency in clause length. In contrast, SCP-AS, which measures sentence complexity, showed a mean of 1.89 and a smaller standard deviation of 0.41, suggesting relatively consistent use of subordinate clauses. Proficiency levels ranged from 5.0 to 7.5, with a mean of 6.4 and a standard deviation of 8.85. This highlights significant variability in participants' language proficiency within the sample.

4.2. RQ1: Correlation with Subjective Ratings

To explore the relationship between objective measures of syntactic complexity (MLAS, MLC, SCP-AS) and subjective ratings of range and accuracy in the TEEP speaking test, Pearson correlation analysis was conducted. The results are shown in Table 6, with effect sizes interpreted based on Plonsky and Oswald's (2014) guidelines: an r value of 0.25 indicates a small effect, 0.40 a medium effect, and 0.60 a large effect.

Table 6. Pearson Correlations.

		Level	**MLAS**	**MLC**	**SCP-AS**
Level	Pearson Correlation	1	.298**	.231*	-.018
	Sig. (2-tailed)		.005	.029	.865
	N	89	89	89	89
MLAS	Pearson Correlation	.298**	1	.284**	.667**
	Sig. (2-tailed)	.005		.007	<.001
	N	89	89	89	89
MLC	Pearson Correlation	.231*	.284**	1	-.427**
	Sig. (2-tailed)	.029	.007		<.001
	N	89	89	89	89
SCP-AS	Pearson Correlation	-.018	.667**	-.427**	1
	Sig. (2-tailed)	.865	<.001	<.001	

N	89	89	89	89

Note: MLAS = Mean Length of AS-unit;MLC=Mean Length of Clause; SCP-AS = Subordinate Clauses Per AS-unit. **Significance levels: *.01*** (2-tailed); *0.05* (2-tailed).

The results revealed a significant positive correlation between MLAS and subjective ratings (*r = 0.298, p = 0.005*), though the small effect size suggests only a slight relationship. Similarly, MLC showed a small but significant correlation with subjective ratings (*r = 0.231, p = 0.029*), indicating that longer clauses were modestly associated with higher proficiency levels.

In contrast, SCP-AS demonstrated no significant correlation with subjective ratings (*r = -0.018, p = 0.865*). This indicates that the use of subordinate clauses was not a meaningful predictor of subjective range and accuracy ratings.

Overall, while MLAS and MLC exhibited slight but statistically significant relationships with subjective ratings, the SCP-AS metric did not contribute meaningfully to subjective assessments. This suggests that other factors, beyond syntactic complexity, play a significant role in influencing subjective ratings.

4.3. RQ2: Differentiation of Proficiency Levels

To address RQ2, several ANOVA tests were conducted to evaluate whether the objective measures of syntactic complexity differed across proficiency levels. Proficiency level served as the independent variable, while syntactic complexity measures (MLAS, MLC, SCP-AS) were the dependent variables. Post-hoc Bonferroni tests were used to identify specific differences between proficiency levels, with a corrected alpha level of 0.017 applied to account for multiple comparisons.

Effect sizes (eta squared) were interpreted using Plonsky and Oswald’s (2014) guidelines: 0.40 for small, 0.70 for medium, and 1.00 for large effects. Descriptive statistics and ANOVA results for each syntactic complexity measure are reported in the following sections.

4.3.1. Mean Length of AS-Unit

The analysis of MLAS showed variability across different proficiency levels. Descriptive statistics (see Table 7) revealed that the overall mean for the sample was 12.28 (SD = 2.35), with scores ranging from 6.87 to 19.20. Specifically, proficiency level 5.0 had the lowest mean of 10.76 (SD = 1.94), while proficiency level 7.5 had the highest mean of 13.44 (SD = 2.47).

Table 7. Descriptive Statistics for MLAS across Proficiency Levels.

N		**Mean**	**Std. Deviation**	**Minimum**	**Maximum**
50	13	10.7608	1.93790	6.87	14.00
55	12	12.1417	2.32045	9.40	16.30
60	9	11.9622	2.64345	9.30	18.00
65	15	12.6133	2.08151	10.40	17.00
70	19	12.0158	2.11247	9.10	16.70
75	21	13.4381	2.46850	10.30	19.20
Total	89	12.2803	2.34848	6.87	19.20

A one-way ANOVA revealed a statistically significant effect of proficiency level on MLAS, with an *F-value of 2.445* and a *p-value of 0.041*. However, given the corrected alpha level of 0.017, this result is not significant, suggesting that MLAS does not significantly vary across proficiency levels. The effect size, measured by eta squared (η^2), was 0.128, indicating a small effect size. This suggests that approximately 12.8% of the variance in MLAS can be attributed to proficiency levels, meaning that proficiency has a small but notable impact on MLAS, though other factors contribute to its variability.

The Bonferroni post-hoc test showed a nearly significant difference between proficiency levels 5.0 and 7.5, with a mean difference of 2.67733 (*p = 0.018*). This suggests that level 7.5 participants produced significantly longer AS-units than those at level 5.0, but no other comparisons were statistically significant.

4.3.2. Mean Length of Clause

Descriptive statistics for MLC showed less variability across proficiency levels, ranging from 6.03 to 6.86, with an overall mean of 6.57. Proficiency level 7.0 had the highest mean at 6.86, and level 6.0 had the lowest mean at 6.03. ANOVA results showed no significant effect of proficiency level on MLC, with an F-value of 1.470 and a p-value of 0.208. This indicates that MLC does not reliably distinguish between proficiency levels in the TEEP speaking test.

4.3.3. Subordinate Clauses per AS-Unit

The descriptive statistics for SCP-AS across proficiency levels revealed scores ranging from 1.7242 (SD = 0.28208) at level 7.0 to 2.0011 (SD = 0.54407) at level 6.0, with an overall mean of 1.8920 (SD = 0.40881). The ANOVA results showed that these differences were not statistically significant, with an F-value of 1.101 and a p-value of 0.366. This suggests that SCP-AS does not reliably differentiate between proficiency levels in the TEEP speaking test.

5. Discussion

5.1. Summaryof Findings

The primary aim of this study was to examine the correlation between objective measures of syntactic complexity MLAS, MLC, and SCP-AS and subjective ratings of grammatical range and accuracy in the TEEP speaking test. The study also sought to determine whether these measures could reliably differentiate proficiency levels. The correlational analysis revealed that MLAS and MLC were positively associated with subjective ratings, though the correlations were weak, indicating that syntactic complexity plays a modest role in shaping ratings of grammatical range and accuracy. When investigating whether these complexity measures could differentiate proficiency levels, only MLAS was found to be a useful measure for distinguishing between proficiency levels in the TEEP speaking test. These results suggest that while syntactic complexity measures provide some insight into proficiency, other factors, such as lexical variety and fluency, likely play a more significant role.

5.2. RQ 1: Correlation between Syntactic Complexity and Subjective Ratings

The study found weak positive correlations between MLAS and MLC and subjective ratings of grammatical range and accuracy. Although statistically significant, the modest strength of these correlations suggests that greater syntactic complexity is only weakly associated with higher ratings. This indicates that, while increasing the length of AS-units and clauses might positively influence ratings, it should be accompanied by improvements in other areas such as lexical variety, fluency, and coherence to more significantly enhance speaking performance.

Interestingly, SCP-AS showed no significant correlation with subjective ratings, which suggests that the use of subordinate clauses may not be a critical factor in assessing grammatical proficiency in the TEEP context. One possible explanation is that raters may not prioritize subordination when assessing range and accuracy. Additionally, factors like lexical variety, fluency, coherence, and pronunciation might have a more substantial impact on the overall ratings, overshadowing the influence of syntactic complexity. Another explanation could be the way complexity and accuracy are assessed together in the TEEP rating scale, which may hinder the clear separation of their individual contributions.

Given these findings, future research should explore the interaction between various linguistic factors, such as lexical variety, fluency, and coherence, to better understand their combined influence on ratings. Additionally, separating complexity and accuracy into distinct criteria could provide a more precise framework for assessing each dimension's contribution to proficiency.

5.3. RQ 2: Differentiation of Proficiency Levels by Syntactic Complexity Measures

This section examines how well objective measures of syntactic complexity MLAS, MLC, and SCP-AS can differentiate between proficiency levels in the TEEP speaking test.

MLAS: The study found no significant differences in MLAS across proficiency levels, which contrasts with some prior studies suggesting that MLAS is a useful indicator of proficiency (Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). While the ANOVA revealed a difference between proficiency levels 5.0 and 7.5, this difference was not statistically significant after correcting for alpha. Previous research has shown MLAS to distinguish between proficiency levels more clearly, particularly between levels 6 and 7 (Bulte & Roothooft, 2020; De Clercq & Housen, 2017). This study's findings suggest that while MLAS may be useful in some contexts, its effectiveness in differentiating closely related proficiency levels is limited.

MLC: The MLC did not show significant differences across proficiency levels in the present study, aligning with findings from other studies (Brown et al., 2008). Previous research has shown a non-linear increase in MLC with proficiency (Bulte & Roothooft, 2020; Iwashita et al., 2008), but this was not observed in this study, likely due to the narrow range of clause lengths. A possible explanation is that at higher proficiency levels, learners may prioritize clarity and coherence over producing longer clauses, as noted by Skehan (2009). This suggests that the MLC measure may not be sensitive enough to capture proficiency differences beyond a certain level.

SCP-AS: The SCP-AS measure also failed to differentiate proficiency levels, a finding that aligns with prior research (Foster & Skehan, 1996). Subordination may not be a reliable marker of proficiency, especially in dialogic speaking tasks, where speakers may prioritize conversational efficiency and turn-taking over syntactic complexity. The nature of the task likely influences the language produced, which might explain the lack of significant variation in SCP-AS scores across proficiency levels.

5.4. Implications of Findings

This study explored two primary research questions regarding the assessment of language proficiency through syntactic complexity measures in the Test of English for Educational Purposes (TEEP) speaking paper. The findings provide valuable insights with significant implications for enhancing the TEEP rating scale, improving task validity, and advancing proficiency assessments in general.

5.4.1. TEEP Rating Scale Enhancement

A key finding of this study was the weak correlation between two syntactic complexity measures Mean Length of AS-unit (MLAS) and Mean Length of Clause (MLC) and subjective ratings. While these measures appear to be modestly associated with higher ratings, their limited impact suggests they are not primary determinants in evaluating proficiency. Furthermore, the Number of Subordinate Clauses per AS-unit (SCP-AS) showed no significant correlation with subjective ratings, which may indicate that evaluators do not prioritize subordination in assessing grammatical proficiency.

This raises the possibility that the TEEP rating scale may benefit from refinement to more distinctly assess complexity and accuracy as separate components. The potential benefits of such a refinement include:

- **Enhanced diagnostic value**, enabling more precise identification of specific language proficiency strengths and weaknesses.
- **Improved fairness** by ensuring that both complexity and accuracy are evaluated independently, preventing one from overshadowing the other.
- **Tailored instructional feedback**, which would offer more actionable insights for students.
- **Increased reliability** in assessment by reducing subjective bias and improving consistency across raters and test administrations.

Refining the TEEP rating scale in this manner would improve the validity of the speaking test and enhance its utility for both assessment and instructional purposes, ultimately leading to more effective language learning outcomes. Future research should further investigate how various linguistic factors, such as fluency, accuracy, lexical variety, and pronunciation, interact to influence ratings in the TEEP context.

5.4.2. Enhancing TEEP Task Validity

The results for RQ2 indicated that while MLAS effectively differentiates high and low proficiency levels, its ability to distinguish between closely related levels is limited. In contrast, MLC and SCP-AS did not demonstrate sufficient sensitivity to proficiency differences, potentially due to the dialogic nature of the task. Given that TEEP already employs a variety of task types, these findings emphasize the importance of designing tasks that elicit a broad range of syntactic structures. Monologic tasks may prompt more complex structures, whereas dialogic tasks may result in shorter, more varied sentences. To more accurately capture test-takers' proficiency, TEEP should ensure that tasks are well-balanced and varied, encouraging the use of a diverse range of syntactic structures.

Future research should examine the impact of different task types (e.g., monologic vs. dialogic) on the elicitation of syntactic complexity. Comparative studies exploring how each task type contributes to proficiency assessments will help refine task design to better capture diverse syntactic structures, enhancing the overall validity of the test.

5.4.3. Advancing Proficiency Assessments

The findings of this study underscore the importance of adopting a multifaceted approach to language proficiency assessment. It is essential to incorporate a variety of syntactic complexity measures, along with other linguistic elements such as lexical and morphological complexity, to provide a more accurate and comprehensive evaluation of proficiency. The research aligns with insights from Bulté and Housen (2012) and Larsen-Freeman (2009), which advocate for multidimensional measures to capture different aspects of linguistic ability.

Moreover, the study highlights the need for task diversity in assessing proficiency. Monologic tasks, such as presentations or extended monologues, may reveal a test-taker's ability to produce complex syntactic structures, while dialogic tasks can better assess interactional competence. The use of a broader array of tasks is crucial in providing a more comprehensive assessment of language proficiency, ultimately enhancing the accuracy and reliability of the evaluation.

6. Conclusion

6.1. Summary of Findings

This study explored the relationship between syntactic complexity measures and subjective ratings in the TEEP speaking test. Key findings include a weak positive correlation between MLAS and MLC with subjective ratings, suggesting that while greater syntactic complexity is weakly associated with higher ratings, it is not a decisive factor. Additionally, SCP-AS showed no significant correlation with ratings, implying limited impact on the evaluation of grammatical proficiency.

The study found that MLAS could differentiate between proficiency levels, particularly between levels 5.0 and 7.5, but did not effectively distinguish among closely related levels. The MLC and SCP-AS lacked sensitivity in differentiating proficiency levels, suggesting their limited applicability in this context.

6.2. Contributions and Implications for Future Research

The study contributes to our understanding of the TEEP speaking test's assessment of syntactic complexity and offers insights into the role of various syntactic measures in differentiating proficiency levels. The findings highlight the need for refining the TEEP rating scale to assess complexity and accuracy separately, which could improve the fairness, consistency, and reliability of the test.

Future research should explore the impact of separating complexity and accuracy in rating scales, investigate the role of additional linguistic factors (e.g., lexical variety, fluency, coherence), and employ mixed-method approaches to provide a richer understanding of language proficiency assessments. Furthermore, comparative studies should examine how different task types, such as monologic versus dialogic, influence the elicitation of syntactic complexity.

6.3. Limitations of the Study

While this study provides valuable insights within the examined context, it is important to acknowledge several limitations related to its design and scope:

1. This study is limited by its narrow focus and context-specific nature, which may restrict the generalizability of the findings to a broader population.
2. The sample size, particularly within each proficiency subgroup, was small, which may affect the stability and representativeness of the results. Future studies should aim to include larger sample sizes to enhance the generalizability of the findings.
3. Additionally, this research focused on a limited set of syntactic complexity measures, and future studies could explore a broader range of fine-grained measures to capture complexity more effectively across different proficiency levels and task types.
4. The study also relied on data derived from a dialogic speaking task, which may differ from tasks that require more formal or prepared speech. Future research could examine a variety of task contexts to provide a more comprehensive understanding of syntactic complexity's role in proficiency assessment.

Acknowledgments: I would like to express my gratitude to the International Study and Language Institute at the University of Reading for their support and for providing the opportunity to undertake this research. I am especially thankful to the Assessment team for granting access to the data that was essential for this study. I am profoundly grateful to my supervisor, Professor Parvaneh Tavakoli, for her expert guidance, constructive feedback, and unwavering support throughout the research process. Her insights and advice have been instrumental in shaping this work.

Conflict of Interest: The author declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bardovi-Harlig, K. (1992). A second look at t-unit analysis: Reconsidering the sentence. *TESOL Quarterly* 26, 390–395.

Biber, D., Gray, B., & Staples, Sh., (2016). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics* 37(5). 639–668.

Bulte, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency. Investigating complexity, accuracy and fluency in SLA* (pp. 21-46). Amsterdam: John Benjamins.

Bulte, B., & Roothooft, H., (2020) Investigating the interrelationship between rated L2 proficiency and linguistic complexity in L2 speech. *Elsevier* https://doi.org/10.1016/j.system.2020.102246

Council of Europe, (2014). *The Common European Framework of Reference for Languages: Learning, teaching, assessment*. https://rm.coe.int/1680459f97

Dörnyei, Z. (2007). *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies*. Oxford University Press

Ehret, K., Berdicevskis, A., Bentz, C., & Blumenthal-Dramé, A. (2023). Measuring language complexity: Challenges and opportunities. *Linguistics Vanguard*, *9*(s1), 1–8. https://doi.org/10.1515/lingvan-2022-0133

Ellis, R. and G. Barkhuizen. 2005. *Analysing learner language.* Oxford: Oxford University Press.

Foster, P., Tonkyn A. and G. Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21: 354–75.

Foster, P., Tonkyn A. and G. Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21: 354–75.

Foster, P., and Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18 (3) (1996), pp. 299-323

Fulcher, G. (2010). *Practical language testing*. London: Hodder Education.

Green, R. (2013). *Statistical analyses for language test developers*. Basingstoke, UK: Palgrave Macmillan.

Harsh, C. (2016). Proficiency. *ELT Journal.* 71(2) DOI:10.1093/elt/ccw067

Housen, A. and F. Kuiken. 2009. Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics* 30(4): 461–473.

Housen, A., & Pierrard, M. (2009). Complexity, Accuracy and Fluency in Second Language Acquisition. *Applied Linguistics*. https://www.academia.edu/28771030/Complexity_Accuracy_and_Fluency_in_Second_Language_Acquisition

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy, and fluency: Definitions, measurements and research. In A. Housen, F. Kuiken & I. Vedder (Eds.), Dimensions of L2 performance and proficiency: complexity, accuracy and fluency in SLA (pp. 1-20). Amsterdam: John Benjamins.

Hulstijn, J. H. (2011) Language Proficiency in Native and Nonnative Speakers: An Agenda for Research and Suggestions for Second-Language Assessment. *Language Assessment Quarterly*. 8:3, 229-249

Iwashita, N., Ortega, L., Rabie, S., & Norris, J. M. (2008). Syntactic complexity and oral proficiency in crosslinguistic perspective. Honolulu, HI: University of Hawai'i National Language Resource Center.

Kang, O., Yan, X., 2018. Linguistic Features Distinguishing Examinees' Speaking Performances at Different Proficiency Levels. *Journal of Language Testing & Assessment.* 1:29-3110.23977/langta.2018.11003

Kuiken, F. (2023). Linguistic complexity in second language acquisition. *Linguistics Vanguard*, *9*(s1), 83–93. https://doi.org/10.1515/lingvan-2021-0112

Kuiken, F., Vedder, I., Housen, A., & De Clercq, B. (2019). Variation in syntactic complexity: Introduction. *International Journal of Applied Linguistics*, *29*(2), 161–170. https://doi.org/10.1111/ijal.12255

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

Norris, J. M. and L. Ortega., (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics* 30(4): 555–578.

Ortega, L. (2003).. Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. *Applied Linguistics.* 24/4: 492-518

Pallotti, G. (2009). CAF: Defining, refining and differentiation constructs. *Applied Linguistics*. 30(4), 590-601. doi:10.1093/applin/amp045

Pallotti, G., (2015). A simple view of linguistic complexity. *Second Language Research*. 31(1) 117–134

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*. 64, 878–912. https://doi.org/10.1111/lang.12079

Seedhouse, P., Harris, A., Naeb, R., & Üstünel, E. (2014). The relationship between speaking features and band descriptors: A mixed method study. *IELTS Research Reports Online Series*, 2, 1e30.

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

Skehan, P. 1998. *A Cognitive Approach to Language Learning*. Oxford University Press.

TEEP Candidate Handbook, (2023). *University of Reading*. https://www.reading.ac.uk/isli/english-language-tests/teep

Vercellotti, M. L. (2015). The Development of Complexity, Accuracy, and Fluency in Second Language Performance: A Longitudinal Study. *Applied Linguistics*. doi: 10.1093/applin/amv002

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity* (Tech. Rep. No. 17). Honolulu: National Foreign Language Resource Center

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.