LEARN Journal: Language Education and Acquisition Research Network
ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 18, No: 1, January – June 2025

Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

Examining Rater Reliability When Using an Analytical Rubric for Oral Presentation Assessments

Sasithorn Limgomolvilas[a], Patsawut Sukserm[b*]

[a] sasithorn.li@chula.ac.th, Chulalongkorn University Language Institute, Thailand
[b] patsawut.s@chula.ac.th, Chulalongkorn University Language Institute, Thailand
[*] Corresponding author, patsawut.s@chula.ac.th

APA Citation:
Limgomolvilas, S. & Sukserm, P. (2025). Examining rater reliability when using an analytical rubric for oral presentation assessments. *LEARN Journal: Language Education and Acquisition Research Network, 18*(1), 110-134. https://doi.org/10.70730/JQGY9980

Received
19/06/2024

Received in revised form
12/08/2024

Accepted
19/08/2024

ABSTRACT

The assessment of English speaking in EFL environments can be inherently subjective and influenced by various factors beyond linguistic ability, including choice of assessment criteria, and even the rubric type. In classroom assessment, the type of rubric recommended for English speaking tasks is the analytical rubric. Driven by three aims, this study analyzes the scores and comments from two raters on 28 video-recorded Thai engineering students' oral presentations using a detailed analytical rubric that covers content, delivery, and visuals. First, it investigates rater reliability by comparing raters' scores using Intraclass Correlation Coefficient (ICC) and ANOVA. Second, applying generalizability theory (G-theory), the correlations between the scores are examined to understand the relationships between different assessment dimensions and how they contribute to a comprehensive evaluation of speaking proficiency. Third, a thematic analysis is performed on raters' comments to gain a deeper understanding of raters' rationale. The findings suggested that a higher number of raters increases the reliability of the ratings, although diminishing returns are observed above a certain threshold. Also, several key themes emerged in relation to the criteria. The study highlights the crucial role of detailed analytical rubrics and cooperation sessions between raters in improving the reliability of oral EFL assessments.

Keywords: analytical rubric, reliability, speaking assessment, generalizability theory, raters

Introduction

Reliability in the assessment of language proficiency is of central importance and forms the basis for the integrity of the evaluation results (Brown, 1995). In the context of assessing EFL learners' speaking skills, reliability ensures that the results are consistent and accurate and reflect the actual abilities of the candidates, regardless of external variables such as different raters or assessment conditions. Among the various reliability measures, inter-rater reliability stands out due to its crucial role in language proficiency assessment, which often involves multiple raters (Brown, 1995; Sundqvist et al., 2020). In fact, when collecting information from raters for assessments, the consistency of scores does in fact need to be examined. Measuring inter-rater reliability can assess the degree of agreement between raters and minimize subjective biases that could distort the evaluation of language performance.

Improving the reliability of speaking performance assessments in EFL contexts has attracted considerable attention in educational research. The search for more reliable assessment methods has led to various strategies aimed at mitigating the challenges of subjectivity and inconsistency in raters' judgments (Burak, 2018; Wind & Peterson, 2017). Apart from the implementation of training programs for raters, another key strategy is the use of detailed analytical rubrics that break down speaking performance into individual components, such as fluency, accuracy, coherence and appropriate use of technical language (Burak, 2018; Davis, 2015; Rubin et al., 1995). These rubrics serve to clarify expectations and provide a structured framework for assessment, thereby reducing the subjectivity associated with the assessment of speaking skills. Research has shown that the reliability of assessments improves significantly when raters use clearly defined rubrics, resulting in a more consistent and equitable assessment of students' speaking skills (Bijani, 2018; Brown, 1995; Huang et al., 2018).

Despite extensive research on inter-rater reliability and the use of analytical rubrics in assessing language proficiency (Brown, 1995; Lumley, 2002), there remains a gap in understanding the correlation between the different assessment dimensions and the specific rationale that raters apply when awarding scores. Previous studies have focused primarily on quantitative measures of reliability without examining the qualitative aspects of rater behavior and decision-making processes (Bachman & Palmer, 2012; Davis, 2015). This study aims to fill this gap by not only examining inter-rater reliability and the effects of varying numbers of raters, but also conducting a thematic analysis of rater comments to uncover the underlying themes and rationales for their rating decisions. By considering both quantitative and qualitative aspects, this research aims to offer a comprehensive understanding of rater reliability and provide practical insights for improving the fairness and consistency of oral presentation ratings in EFL contexts.

The importance of considering the number of raters involved in the assessment process has also been emphasized. Studies based on generalizability theory (G-theory) have investigated how the number of raters affects the reliability of speaking ratings and have found that a higher number of raters can improve the reliability of ratings (Burak, 2018; Fan & Yan, 2020; Hidri, 2018; Tran & Hang, 2021). Nevertheless, practical limitations such as the availability of trained raters (Fan & Yan, 2020) should also be considered. Limited resources and time mean that recruiting multiple raters for speaking assessments is often infeasible, especially if these raters are also expected to provide feedback to individual students (Putri et al., 2019).

This study examines the use of a detailed analytical rubric in an oral presentation course for Thai engineering students. It is driven by three main objectives. First, it aims to investigate the inter-rater reliability in the assessment of these students' speaking performance. Second, the study draws on generalizability theory to examine the effects of varying numbers of raters. Third, the study examines the raters' reasoning process when allocating scores to English oral presentation assessments. These aims are important in that they address key challenges in the assessment of speaking skills and aim to improve the reliability of such assessments (Bijani, 2018; Brown, 1995; Sundqvist et al., 2020). In this way, the study can help to ensure that EFL students are assessed in a way that truly reflects their abilities and potential.

Research Questions

1. What is the inter-rater reliability in the assessment of students' speaking performance?
2. How does the varying number of raters affect the assessment of students' speaking performance, as examined by generalizability theory?
3. What is the raters' reasoning process when allocating scores to English oral presentation assessments?

Literature Review

The assessment of speaking performance in EFL environments involves a variety of methods, each of which has its own advantages and challenges (Putri et al., 2019; Tran & Hang, 2021). Traditional assessment methods range from face-to-face oral interviews to group discussions and presentations (Huang et al., 2018). More modern approaches include the use of digital recordings and computer-assisted communication, which offer new opportunities for the assessment of speaking skills (Gan, 2013). These methods aim to measure a range of language competencies, including fluency, coherence, grammatical accuracy, and appropriate use of technical vocabulary (Lee, 2007; Ugiljon, 2018).

Despite the variety of assessment methods, the literature repeatedly points to a critical problem: the limitations of accurate assessment of speaking performance (Iberri-Shea & Hui, 2017), one of which is the issue of reliability. When evaluating reliability, four strategies can be applied: test-retest reliability, parallel-form reliability, internal consistency, and marker reliability. In speaking assessment, where the evaluation process relies chiefly on the rater's judgement, marker reliability is the focal point to achieve. Two aspects of marker reliability are intra-rater and inter-rater reliability. When one rater determines the score, the only thing that matters is the rater's intra-rater consistency. If there are multiple raters, inter-rater reliability is important to maintain uniformity among the raters.

Inter-rater reliability (IRR) is of crucial importance in studies that involve subjective measurements, particularly the assessment of speaking performance (Bruton et al., 2000). It assesses the extent to which different raters assess the same phenomenon consistently and thus provides a measure of the reliability of the ratings. High inter-rater reliability indicates that the assessment method provides stable and consistent results, regardless of who is carrying out the assessment (Stolarova et al., 2014).

The subjective nature of language assessments can lead to significant variations in the scores given by different raters based on their perceptions, biases and interpretations of the criteria (Ekmekçi, 2016; Lee, 2007; Wind & Peterson, 2017). EFL environments present particular challenges for maintaining high inter-rater reliability. These include the different linguistic backgrounds of raters, varying degrees of familiarity with the specific jargon of academic disciplines, and differences in personal evaluation style (Brown, 1995; Lamprianou et al., 2021; Wind & Peterson, 2017). Research has consistently shown that variability in raters' judgments can significantly affect the results of speaking assessments (Huang et al., 2018). Studies have also shown that without structured training and clear scoring rubrics, raters may apply the criteria inconsistently, leading to unreliable scoring results (Bijani, 2018; Brown, 1995; Burak, 2018). It is thus crucial for raters to understand the rubric and conform with the standard.

To address this issue, research suggests several strategies, including the development of detailed, discipline-specific rubrics and comprehensive training programs for raters. These approaches aim to improve inter-rater reliability by clarifying scoring criteria and ensuring that raters have a shared understanding of the constructs being assessed (Brown, 1995; Lumley, 2002; Sundqvist et al., 2020). In addition, involving multiple raters and using statistical methods to measure reliability are also recommended practices to mitigate the subjective nature of speaking ratings (Burak, 2018; Ekmekçi, 2016; Sundqvist et al., 2020). Ultimately, improving the reliability of assessments of EFL students' speaking performance requires a concerted effort to address the inherent challenges of subjective assessment, which underscores the need for continued research and innovation in assessment practices (Bijani, 2018; Burak, 2018; Davis, 2015; Sundqvist et al., 2020).

Generalizability Theory

Generalizability Theory, developed by Lee Joseph Cronbach and his colleagues, provides a comprehensive framework for evaluating the reliability of psychological tests, educational assessments, and various other types of measurements (Leung, 2015). G-Theory offers a methodological approach to disentangle the multiple sources of error that affect measurement and provides a more sophisticated analysis than classical test theory, which generally considers only one source of error at a time (Brennan, 2001).

In educational contexts, G-Theory is helpful in evaluating the reliability of assessments by analyzing the extent to which different sources of variability (e.g., assessors, tasks, occasions) influence the results. This analysis helps understand how different criteria of an assessment contribute to overall measurement error, thus providing educators with guidance on how to improve the reliability of their assessments (Vacha-Haase et al., 1998). An analysis based on G-theory usually consists of two phases: a G-study and a D-study. In a G-study (generalizability study), researchers estimate the various sources of error and their extent. The subsequent D-study (decision study) uses this information to develop the most efficient and reliable measurement procedures tailored to specific purposes (Shavelson & Webb, 1991).

For the assessment of speaking proficiency, especially in settings where English is a second language, G-theory provides a valuable framework for researchers and educators to examine how various factors, such as rater bias, task difficulty, and test conditions, affect the consistency of speaking scores (Jason & Xun, 2020). Although it is not the only method available, studies using G-Theory in EFL settings have highlighted its usefulness in improving the fairness and accuracy of speaking proficiency assessments (Lee, 2007). By identifying key sources of error, institutions can implement more reliable assessment strategies, such as standardized assessment protocols, to ensure that results more accurately reflect students' actual language proficiency.

In the context of assessing EFL students' speaking performance, the application of G-Theory would enable this study to systematically investigate and address the variability in raters' scores to improve the reliability and fairness of the assessments. By examining how the varying number of raters affects the generalizability of the ratings, the study aims to create more robust and equitable assessment frameworks.

Methodology

This research integrates quantitative and qualitative data to provide a deeper insight into assessment reliability. The choice of this design was justified by the objectives of the study, which require an examination of numerical data to assess inter-rater reliability, as well as an exploration of raters' explanations.

A generalizability study (G-study) was also designed to investigate the sources of measurement error that might affect the reliability of EFL students' oral presentation ratings. In this study, the "individual sources" refer to the various components that contribute to the overall variance in the assessment results. Specifically, these are the following sources.

Table 1

Components of Source Variances in Rater Assessment

Sources	Definition
Student (S)	The individual differences between the students being assessed.
Rater (R)	The individual differences between the raters scoring the presentations.
Criteria (C)	The specific dimensions or criteria of the rubric used for scoring, which are content, delivery, visuals.
Interactions	The combined effects of these components, such as how different raters may score different criteria for the same student e.g., student-rater interaction (SR), student-criteria interaction (SC), rater-criteria interaction (RC), and student-rater-criteria interaction (SRC).

To quantify the amount of variance, analysis of variance (ANOVA) methods, which allow the partitioning of the observed variance into its individual sources, were used. These results were crucial for the design of the subsequent decision study (D-study), which used this information to optimize the assessment design and improve the reliability of our assessment procedure.

That is following the G-study, a decision study (D-study) was employed for the identified variance components to predict the reliability of different assessment designs. This study examined how the reliability of the ratings could be optimized by varying the number of raters. To this end, the generalizability coefficient (G coefficient) was calculated for different numbers of raters to determine the ideal number that reconciled reliability with practical constraints such as the availability of raters. The D-study aimed to help make informed decisions about the optimal assessment conditions that minimize error while ensuring efficient use of resources.

Meanwhile, qualitative analysis was conducted on raters' comments to understand the rationales that may have influenced the reliability of the speaking assessments. These included perceptions of scoring criteria, and the challenges faced when assessing EFL students' oral presentation.

Participants

The main participants were two raters who assessed an intact group of 28 engineering EFL students from a public university who were enrolled in a listening and speaking course, Communication and Presentation Skills, in the academic year 2021. The two selected raters are native speakers who have more than five years of experience in teaching this course. They were the main instructors who helped develop the course and the task. Prior to and after scoring, the raters had a brief training session to familiarize themselves with the analytical rubric and calibrate their scoring. EFL students were from mixed majors of computer engineering, industrial engineering and nuclear engineering. These engineering students were mostly in their second year with a few of them in their third year.

Rater Selection

The credibility of the research results depends to a large extent on the competence and consistency of the raters. The raters were selected based on several criteria. Firstly, both raters hold a degree or are certified in English language education. Their academic qualifications provide a solid foundation in teaching and assessment skills. Secondly, the selected raters have more than five years of teaching experience in this course. Their extensive experience in teaching and assessing speaking skills ensures that they are familiar with the course content and expectations for student performance. Thirdly, a consultation session after the assessment was conducted every semester to align their scoring standards. The consultation between raters aimed to minimize rater bias and ensure consistent application of the rubric. Finally, raters have proven to be reliable on previous assessments. Their past performance, as evidenced by high inter-rater reliability scores in a similar study (Naphon, 2017), demonstrates that they can consistently and accurately rate speaking performance. Therefore, the selection of these raters was based on their qualifications, experience and proven reliability, which are critical to the credibility of the research findings.

Oral presentation task

As part of their course, students completed an oral presentation, which accounted for 15% of the grade. The task was designed to assess the ability to orally present technical content on an engineering-related topic to a non-technical audience with a time limitation between three and four minutes. Since the class was delivered during the COVID pandemic, students were required to complete their presentation as a video recording. Students were not permitted to edit the audio components of their video, but were permitted to edit the visual elements to incorporate their slides. Students were assigned the task, and given a month to prepare and practice the presentation with their instructor before submitting their video recording. A Google drive link for each class was provided for students to upload their materials as well as the presentation slides. Two raters then independently evaluated the students' presentations based on the same rubric (see Appendix) and provided explanations for the scores on an Excel sheet organized by the course coordinator. Before the pandemic, a cooperation session between raters after assessment was held immediately after the presentation for raters to discuss scores agreement as well as their comments. During the pandemic, raters discussed with each other only when the score difference was over three.

Rubric for Oral Presentation

The rubric for the oral presentation was a modified version of a rubric that had been developed several years earlier, which was reported by Naphon (2017) to exhibit strong inter-rater reliability. This rubric was chosen because it has proven to be effective in the uniform and fair assessment of oral presentations. It breaks down the scoring into specific, detailed components and ensures that scorers have clear criteria to follow, reducing subjectivity and increasing reliability. The current version of this rubric comprises three specific dimensions of oral presentations: Content (Story and Organization), Delivery (Body language and Language), and Visuals. For each dimension, there are detailed descriptors with examples of five different levels of performance (fail, poor, average, strong, and superior) allowing a range of scores from 1 to 10. In terms of Content, this area evaluates the organization of the presentation, the clarity of the main idea, the logical flow of information, and the ability to engage the audience with a coherent and compelling narrative. It includes criteria such as the effectiveness of the introduction, the development of the main points, the use of evidence and examples, and the strength of the conclusion. Regarding Delivery, this section assesses the speaker's body language, including eye contact, gestures and posture, as well as vocal characteristics such as volume, clarity, pace and intonation. It also examines the flow of speech, pronunciation and use of language, including grammar and vocabulary. For Visuals, this section assesses the quality and effectiveness of the visual aids used in the presentation. This includes the design and clarity of the slides, the appropriateness and integration of images and text, and how well the visual aids support and enhance the verbal presentation. The rubric was distributed to students when assigning the task.

Data Collection

The scores and comments were collected from the two qualified raters for an intact group of 28 students, out of a total cohort of 12 sections. The quantitative data included the numerical scores for each dimension, while the qualitative data consisted of the raters' written comments explaining their scoring decisions. In other words, the quantitative data consisted of the scores given by two raters for each of the 28 oral presentations given by the students. Each oral presentation comprises five sets of data, which are scores on content, scores on delivery, scores on visuals, total scores, and comments. The qualitative data consisted of the written comments made by the raters during the evaluation of the oral presentations. These comments were recorded in an Excel spreadsheet organized by the course coordinator. The comments were then analyzed thematically to identify key themes and patterns in the rater's behavior and decision making.

Data Analysis

The quantitative data obtained from the raters' ratings were analyzed using statistical software. The analysis focused on the calculation of the inter-rater reliability coefficient. Inter-rater reliability was assessed using intraclass correlation coefficients (ICC) for continuous rating scales, which are a measure of the degree of agreement between raters. Moreover, the effect of different numbers of raters on the reliability of scores was investigated using generalizability theory (G-theory). The data analysis comprised two main phases: a Generalizability study (G-study) and a decision study (D-study). First, the G-study was conducted to identify and estimate the various sources of error in the assessment process. For this purpose, the total variance of the assessment results was broken down into components attributable to the raters, the criteria assessed and the students, as well as their interactions. Using analysis of variance (ANOVA), we were able to partition the observed variance into these individual sources. This step helped understand how much each factor contributed to the overall measurement error.

Next, the D-study used the information from the G-study to predict the reliability of different assessment designs. In this phase, how increasing the number of raters would affect the reliability of the results was examined. The generalizability coefficient (G coefficient) for various numbers of raters was calculated to find the optimal balance between reliability and practical constraints. By varying the number of raters, it is clearly seen to what extent reliability may be improved and at which point adding more raters would provide diminishing returns. These two phases allowed us to comprehensively analyze and optimize the assessment process to ensure that the ratings were both reliable and practical to implement. Qualitative data were also examined through a thematic analysis of raters' comments to gain a deeper understanding of the criteria and rationale used during the assessment process.

Findings

The main findings of this study show that the inter-rater reliability of the assessments was generally moderate to good, and that higher reliability would be achieved by using multiple raters. The analysis revealed significant correlations between the different assessment dimensions, underlining the importance of a detailed analytical rubric. In addition, thematic analysis of the raters' comments identified several key themes that influenced rating decisions, including Content, Delivery and Visuals. These findings highlight the effectiveness of structured rubrics and the potential benefits of rater collaboration in improving the consistency of oral presentation assessments.

Reliability in the Assessment of EFL Students' Speaking Performance among Raters

Table 2 below shows the scoring of two different raters evaluating criteria of EFL students' oral presentations. These criteria were divided into three categories: Content, Delivery and Visuals.

Table 2

Raters' Scores of English Oral Presentations

Criteria	Rater 1		Rater 2	
	M	***SD***	***M***	***SD***
Content	7.66	0.96	7.23	1.40
Delivery	6.92	1.35	7.44	1.23
Visuals	7.41	0.77	7.74	0.62
Total	21.60	2.69	22.17	2.57

The ratings of the two raters appear to be quite similar, with only minor differences in the various criteria. Rater 1's rating had slightly lower mean scores for delivery (*M* = 6.92*; SD* = 1.35) and overall rating (*M* = 21.60*; SD* = 2.69) compared to Rater 2 (*M* = 22.17*; SD* = 2.57). These differences indicate that Rater 1 might apply a stricter interpretation of the delivery criterion, which was reflected in a slightly lower mean and a higher standard deviation. However, both raters gave relatively similar scores, indicating a general consistency in their ratings.

Table 3

Intraclass Correlation Coefficient

Types	Intraclass Correlation	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.575	.263	.778	3.706	27	27	.001*
Average Measures	.730	.417	.875	3.706	27	27	.001*

*p < .05

To verify reliability, an inter-rater reliability check was conducted. Table 3 shows two types of intraclass correlation coefficients (ICCs)—single measures and average measures—with the corresponding 95% confidence intervals and significance levels of an F-test. The ICC for individual measure was .575, indicating moderate agreement between individual raters. The 95% confidence interval ranged from .263 to .778, indicating that the true ICC may lie within this range. The F-test value of 3.706 with 27 degrees of freedom for both the numerator (df1) and the denominator (df2) was statistically significant (p = .001), meaning that the agreement between raters was significantly better than would be expected by chance.

In terms of average measures ICC, the value was .730, which is generally considered a good level of reliability and indicates that the average ratings of the raters were consistent with each other. The 95% confidence interval for the ICC of the average scores was wider, ranging from .417 to .875, indicating stronger agreement between raters when their ratings were averaged. The same F-test statistics were found as for the individual measures, indicating statistical significance.

According to Koo and Li (2016), ICC values between 0.5 and 0.75 indicate moderate reliability, while values between 0.75 and 0.9 indicate good reliability. The results, therefore, demonstrate moderate to good agreement between raters in their ratings of oral presentations, and this agreement was statistically significant. The average scores provided a more reliable rating than the individual scores, which is typical as averaging generally reduces random error.

The Effects of Varying Numbers of Raters

Table 4

G-study for S x R x C Design Where Each Student (S) was Assessed in Three Criteria (C) by Two Raters (R)

Source of variance	Sum of Squares	df	Mean Square	Estimated Error components	Percentage of Error variance
S	123.78571	27	4.58466	.71017	37.5
R	0.59524	1	0.59524	.00323	0.2
C	17.19048	2	8.59524	.07349	3.9
SR	8.73810	27	0.32363	.10788	5.7
SC	70.14286	54	1.29894	.34215	18.1
RC	3.47619	2	1.73810	.04012	2.1
SRC, e	33.19048	54	0.61464	.61464	32.5
Total	257.11905	167			100

Table 4 contains the results of a G-study investigating the sources of variance in the assessment of oral presentations, examining how the students (S), raters (R) and criteria (C) of the presentations (content, delivery, and visuals) contributed to the variance in the overall evaluation as follows:

S: Variance among the students accounted for 37.50% of the total variance, with an estimated variance component of 0.71, the largest contribution, suggesting that different students had a significant impact on scores.

R: The variance attributable to raters was almost negligible, with an estimated component of <.01 and only 0.20% of the total variance, indicating that rater bias did not have a significant impact on assessment outcomes.

C: The variance due to different criteria of the presentations considered in the assessment process contributed to 3.90% of the total variance with an estimated component of 0.07, suggesting that it had some influence on the scores but was not as significant as the sentences.

SR interaction: The interaction between students and raters had a small, estimated variance component of 0.11 and accounts for 5.70% of the total variance.

SC interaction: This interaction was more pronounced, with a variance component of 0.34 and accounting for 18.10% of the total variance, suggesting that the way different criteria were rated could vary considerably depending on the students.

RC interaction: With a component of 0.04, the interaction between raters and criteria contributed only slightly (2.10%) to the total variance.

SRC interaction: This triple interaction had a considerable influence with an estimated variance component of 0.61 and a share of 32.50% of the total variance. This means that the combination of the factor's student, rater and criteria introduced a considerable degree of variability into the ratings.

Overall, the study showed that the largest sources of variance were the students who completed the assessment and the complex interactions between students, raters and criteria, which together accounted for a significant proportion of the total variance in the assessment. Variability due to raters alone was minimal, indicating consistent scoring by different raters. However, when raters interacted with different students and criteria, variability increased, which could indicate that certain raters rated some criteria differently depending on the specific students who completed the scenario.

Table 5

D-Study for S x R x C When Number of Criteria in Rubric = 3 and Number of Raters Ranging from 1 to 5

	Rater (s)	***1***	***2***	***3***	***4***	***5***
	Criteria	***3***	***3***	***3***	***3***	***3***
Error Variance	σ^2_{Rel}	0.10788	0.05394	0.03596	0.02697	0.02158
	σ^2_{Abs}	0.11111	0.05556	0.03704	0.02778	0.02222
G-Coefficient	ρ^2_{Rel}	0.86813	0.92941	0.95181	0.96341	0.97051
	ρ^2_{Abs}	0.86471	0.92745	0.95043	0.96236	0.96966

Table 5 shows the results of a D-study, which was an extension of the theory of generalizability. This study examined how different numbers of raters would affect the reliability of scores for oral presentations assessed under three criteria. Firstly, relative error variance (σ^2_{Rel}) decreased as the number of raters increased. With one rater, the error variance is 0.10788, but it decreased continuously to 0.02158 when there were five raters. Meanwhile, absolute error variance (σ^2_{Abs}) decreased as the number of raters increased, from 0.11111 for one rater to 0.02222 for five raters.

The G-coefficient for relative generalizability (ρ^2_{Rel}) increased with the number of raters. It started at 0.86813 for one rater and improved to 0.97051 for five raters. In the same way, the coefficient for absolute generalizability (ρ^2_{Abs}) reflected the reliability of the absolute decisions. It followed a similar pattern to the relative coefficient and increased from 0.86471 for one rater to 0.96966 for five raters.

Therefore, the results of the D-study clearly show that the reliability of the oral presentation assessment improved significantly as the number of raters increased. When the number of raters increased from one to five, both the relative and the absolute G coefficient approached the value of 1, which indicated a very high reliability. Nevertheless, the aim for these coefficients was to reach or exceed a threshold of 0.80, which is generally considered the minimum for an assessment to be deemed highly reliable. That is, a G-coefficient of 0.86813 or 0.86471 for a single rater was within this acceptable range, particularly when the addition of more raters may not be feasible due to resource constraints. As Crocker and Algina (1986) point out, in cases where achieving perfect reliability is challenging—such as with subjective tasks like oral presentations—a reliability coefficient above 0.80 is often sufficient to ensure confidence in the results. Thus, the G-coefficients for both relative and absolute reliability, which were slightly above this threshold, indicate that a single rater is likely adequate for reliable scoring of oral presentations across the three criteria. Furthermore, Nunnally (1978) claims that a reliability coefficient of 0.70 or more may be acceptable in the initial stages of research, while higher values — such as 0.80 or more — are preferable in applied settings where the ratings influence important decisions. Therefore, the single-rater reliability observed here, which was just above the 0.80 mark, supports the argument that the reliability of the assessment was sufficient for practical purposes.

Reasons for the Assigned Scores

In analyzing the raters' comments on the oral presentation assessments, several key themes emerged in relation to the three main criteria in the rubric: "Presentation not bad, but voice a bit soft; lacks energy/enthusiasm. Memorized script delivered in monotone voice. Some mispronunciations. Pretty good PowerPoint. Well organized and interesting content.” This comment is one of the comments from the raters. Although different themes were found, they had been classified here for better comprehension according to the rubric. Subcategories were identified and categorized under the three criteria. These themes help understand what influences the raters’ scoring decisions.

Theme 1: Delivery

After dividing comments by categories and tone (positive or negative), the comments were further analyzed in comparison with the detail in the rubric as shown in Table 6. As a result, raters’ comments related to the students’ delivery can be sorted into three subcategories: language, body language and voice. The comments on language mentioned language and pronunciation errors, pauses and pace. These terms can be referred to as language fluency. In fact, when students could display good language, pronunciation, and pace, they received a comment of “*excellent fluency*” or *“good fluency”*. Second, body language involved gestures, postures and eye contact. Not many comments on body language were identified since many of the students recorded their video while sitting. Most of the comments in this theme related to eye contact and movement.

The third sub-theme, voice, entailed intonation, volume, and naturalness. This indicates another fluency aspect as proficient English users should know how to put stress on words and vary intonation in sentences. At times, volume presented a problem to the raters’ understanding as some speakers spoke too quietly. In addition, a theme of students reading from a script was detected in several subcategories with rater’s experience. This suggests that many students did not put as much effort into their performance they were expected to due to the convenience of technology.

Table 6

Themes on Delivery

Delivery (Body language & Language)				
Voice	Good voice control	5	Flat/monotone	12
	Okay voice control	3	Scripted	6
	Enthusiastic	3	Unenthusiastic/Uninterested	4
	Loud voice	1	No vocal variety	3
			Almost too low/ Hard to hear	3
			Seemed dubbed	2
			Exaggerated	1

			Seemed to go up and down	1
			Issue with intonation	1
Pace	Good fluency	5	Reading from script	7
	Okay fluency	3	Pauses	4
	Excellent fluency	1	Struggled	3
	Fairly fluent	1	Fluency issues	1
			Pace a bit quick	2
Body	Good/nice use of hand gestures	4	No eye contact	5
	Good body language	2	Eye contact fixed	4
	Some gestures	1	Can't assess gestures/ No gestures	3
			Body lang a bit nervous	1
			Eye control all over	1
			Eyes moving like reading a script	1
Language	Excellent English	1	Some pronunciation and lang errors	9
			Some minor language and pronunciation errors/ issues	9
			Not good/ Poor enunciation/ No energy	8
			Grammar error	2
			Some target language errors	1
Overall	Good delivery	6	Looking at notes a bit	1
	Confident	2	Negative delivery	1
	Average delivery/ presentation	2	Sluggish delivery	1
	Strong delivery	1		
	Good energy	1		
	Natural Delivery	1		
	Genuine effort on delivery with the eyes, vocal variety, smiles	1		

Theme 2: Content and Organization

Raters paid close attention to how well presenters caught the audience’s attention and explained their content, particularly in relation to technical terms. The subcategories standing out under this theme were hook, structure and transitions, conclusion, and content. It is worth noticing that one rater focused consistently on the hook and transitions, making both positive and negative comments, while the other rater made very few positive comments on these criteria. This reflects the nature of the first rater who tended to explain both the positive and negative qualities. The second rater, however, was inclined to only comment on certain criteria if the student did not meet a high standard. On some occasions, the second rater even provided a long explanation on how bad the presentation was.

Apart from that, both raters appreciated presenters who could convey complex concepts in understandable terms, suggesting that the ability to translate technical information for a general audience was an important criterion. Accordingly, three other common themes in raters' comments on content, which were somewhat intertwined, were length, technical content, and interestingness. The integration of technical or engineering content was one way for students to prove their ability in presenting their background knowledge to a non-technical audience in an interesting way and a specified time.

Theme 3: Visuals

Four subcategories under this theme were overall, text, image, and presentation. The main word used to describe overall visuals was *okay*, while other common comments involved the issue of inconsistency and clutter. Text and image were the only two categories that did not contain any positive words. In addition, *wordy* was found to be the major negative comment for text. Presentation was mentioned by both raters in terms of the presenter's effectiveness and delivery when using PowerPoint or slides. A few comments described how presenters could not coordinate the visuals to maintain the audience's attention.

Theme 4: Others

Apart from the themes and subcategories above, other keywords related to time and professionalism. They are listed in the rubric as requirements that students must follow to avoid score penalties. For example, scores were reduced for any presentation above or under the time limit of three to four minutes. Professionalism in this case refers to the dress code, which required students to wear a proper uniform. Time and professionalism did not fall under any categories above and these penalties were deducted from the total scores.

Table 7

Themes (Categorized into Subcategories)

Content	Delivery	Visuals
-Hook -Structure and transition -Content -Conclusion	-Language -Body language -Voice	-Overall -Text -Image -Presentation

In summary, raters' comments demonstrated a multi-faceted approach to evaluating oral presentations, assessing not only the content but also the way in which the presenter delivered and presented it with visuals. As shown in Table 7, raters emphasized the importance of structuring the technical content to captivate the audience's attention, dynamically supporting the content with visuals, and using language skills effectively.

Discussion

Inter-rater Consistency

The investigation of inter-rater reliability (IRR) methods in the assessment of students' speaking performance has revealed several insightful findings that are consistent with previous studies while providing a nuanced understanding of the challenges and strategies for ensuring and improving inter-rater consistency.

The study underlines the high degree of reliability in raters' assessments, which can be attributed to the standardized training of the raters and the use of a detailed analytical rubric. This positive result suggests that the structured training significantly improves the consistency of judgments by ensuring a standardized interpretation of the evaluation criteria, as also demonstrated by the work of Burak (2018) and Brown (1995). The detailed scoring rubrics effectively guide raters' judgments and increase the likelihood that all criteria of speaking performance are scored consistently and reliably (Rubin et al., 1995; Wind & Peterson, 2017). These procedures have resulted in high interrater reliability, demonstrating the effectiveness of our scoring procedures. In this study, the detailed descriptors in the analytical rubric may have helped raters to conform to scores within an agreeable range.

The G-study found that while the variability attributable to raters alone was minimal, suggesting basic consistency between raters, the interaction effects between raters and students, and between raters and criteria, resulted in greater variability. This suggests that while raters were generally consistent in their ratings, their interpretations of different criteria could vary, especially when interacting with different student performances. This finding is consistent with studies that emphasize the complexity of language assessments and the need for extensive rater training that considers not only consistency of scoring but also the application of criteria in different contexts (Brown, 1995; Lumley, 2002).

The significant contribution of student-aspect interaction to the total variance emphasizes the multidimensional nature of assessment of speaking performance. It reflects the challenges of evaluating performance in different content areas and criteria, which can vary considerably from student to student. This is particularly critical in EFL education, where the ability to communicate content clearly and effectively is as important as verbal fluency (Davis, 2010).

Effect of Numbers of Raters on Assessment of Speaking Performance

The results of the application of generalizability theory suggest that the inclusion of a larger number of raters could improve the reliability of ratings. This underpins the research findings of Bijani (2018) and Sundqvist et al. (2020), who assume that a higher number of raters can compensate for individual biases and thus improve the objectivity of the results. The D-study extended these findings by quantitatively analyzing how increasing the number of raters would affect the reliability of the rating results. The results showed a clear trend: as the number of raters increased, the error variance decreased, leading to an increase in reliability estimates. This supports the hypothesis that the use of multiple raters can help mitigate the bias of a single rater and lead to a more reliable rating, as suggested by Hidri (2018) and Tran and Hang (2021).

However, it is noteworthy that, given the generalization coefficients, the reliability met an acceptable threshold for a single rater. According to Nunnally (1978), a reliability coefficient of 0.70 or more may be acceptable in the initial stages of research, but higher values — such as 0.80 or more — are preferable in an applied setting where decisions are based on the ratings. This indicates that a single rater with sufficient experience and training could achieve a level of reliability that would be generally considered satisfactory for practical purposes. Crocker and Algina (1986) also emphasized that in situations where perfect reliability is difficult to achieve, particularly in subjective tasks such as oral presentations, a coefficient above 0.80 is often sufficient to ensure confidence in the results. While adding more raters did increase reliability, particularly beyond three raters, the gains were not as significant. These results show that although multiple raters can enhance reliability, a single, well-trained rater was often able to provide reliable assessments, especially in situations where resources are limited. It suggests that, with detailed analytical rubric and appropriate training session, the efficiency of the assessment process can be maintained without compromising reliability, even if only one assessor is used.

Themes of Reasoning in English Oral Presentation Assessment

The thematic analysis of raters' justifications for the evaluation of EFL students' oral presentations provides a nuanced understanding of the criteria that are considered important in the evaluation process and how raters' interpretations align with the established rubric. The analysis revealed several main themes that influenced raters' scoring decisions: voice control in speech delivery, professionalism, visual presentation, content understanding, structure and explanation, and technical and linguistic competence. These themes reflect the complexity of effective communication in technical fields as described in the literature (Lee, 2007; Lumley, 2002).

Raters frequently commented on the importance of voice control, emphasizing factors such as monotone and enthusiasm, which is consistent with Rubin et al.'s (1995) findings that effective communication requires not only clarity but also engagement through voice modulation. It underlines that a speaker's ability to vary their voice in delivery can significantly affect their clarity and engagement with audiences.

Furthermore, the quality of visual aids reflects the broader academic and professional standards expected in technical contexts (Davis, 2010). The visual aids from this study were assessed in several categories in terms of whether they match professional presentation etiquette. This theme aligns with studies such as that of Kaewpet and Sukamolson (2011), which emphasize the importance of nonverbal cues in professional communication.

Raters valued presenters' deep understanding of content and ability to clearly explain complex technical concepts with clear structure. As Drubin and Kellogg (2012) noted, these skills are critical in technical disciplines, as the ability to articulate complex ideas concisely is essential for academic and professional success. Finally, technical accuracy and linguistic accuracy were of critical importance to raters, reflecting the dual requirements of language skills and technical expertise in engineering education, as discussed in the literature by Orr (2010).

The identification of these themes highlights the complexity of assessing oral presentations in EFL contexts. The findings suggest that raters place great importance not only on the linguistic criteria of the presentation, but also on other criteria including nonverbal communication, visuals and technical content. This is in conformity with the findings of Lumley (2002) and Rubin et al. (1995), who argue for the development of detailed rubrics that can guide raters' judgments more effectively. Nonetheless, this complexity can pose a challenge to raters, who need to assess so many variables at the same time and within a time limit.

Conclusion and Implications

The study highlights how the creation of a detailed analytical rubric can mitigate the considerable variability in raters' judgments. The variability arises not only from the subjectivity of perception, but also from inconsistent understanding and application of assessment criteria. Cooperation sessions between raters after assessment and the use of detailed rubrics can therefore improve the consistency of raters, as shown by the intraclass correlation coefficients and qualitative thematic analysis of rater comments.

The results of the study, based on generalizability theory, show that a higher number of raters generally increases the reliability of ratings. However, there is a point of diminishing returns where additional raters no longer significantly improve reliability. In fact, reliability reached an acceptable threshold, suggesting a balance between reliable ratings and maintaining the efficiency of the assessment process with two raters or even one rater. This balance would be particularly important in educational settings where resources and time are limited. Therefore, the use of a rater can be justified not only out of necessity, but also because it meets the requirements of reliability for practical assessment purposes, as demonstrated by the results of this study.

Through a thematic analysis of raters' comments, the study provides insights into the criteria that influence scoring decisions, such as delivery style, content explanation and visuals. Although the words that the two raters commented were slightly different, the themes were largely similar. This could stem from the raters' cooperation sessions and the development of clear, standardized analytical rubrics that were found to be effective in reducing variability and increasing the reliability of ratings. These common themes could act as guidance in the development of formative assessment that can aid and save time for raters and learners in the learning process before the summative assessment. This comprehensive approach would be essential to accurately assess the communicative competence of EFL students when performing oral presentations in their field of study.

About the Authors

Sasithorn Limgomolvilas: A full-time lecturer at Chulalongkorn University Language Institute (CULI). She has a Bachelor's degree in Education from Chulalongkorn University, and a Master's degree in Teaching English as a Speaker of Other Languages (TESOL) from San Francisco State University. She later taught at CULI and earned a Ph.D. in English as an International Language (EIL) from Chulalongkorn University. Her primary areas of research focus are on the fields of ESP and language assessment.

Patsawut Sukserm: A full time English lecturer at Chulalongkorn University Language Institute, Thailand. He holds a Ph.D. in English as an International Language Program, Graduate School, Chulalongkorn University. He also has a B.A. English (1st class honors) from Ramkhamhaeng University, a B.S. in Statistics and an M.A. in English as an International Language from

Chulalongkorn University. His areas of research include quantitative research, language testing and assessment, and English language teaching.

References

Bachman, L. F., & Palmer, A. S. (2012). *Language assessment in practice*. Oxford University Press.

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education, 5*(1), 1-20. https://doi.org/10.1080/2331186x.2018.1460901

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15. https://doi.org/10.1177/026553229501200101

Bruton, A., Conway, J. H., & Holgate, S. T. (2000). Reliability: What is it, and how is it measured? *Physiotherapy, 86*(2), 94-99. https://doi.org/10.1016/S0031-9406(05)61211-4

Burak, M. (2018). Speaking assessment: Impact of training sessions. *World Science, 2*(12(40)), 44-48. https://doi.org/10.31435/rsglobal_ws/30122018/6275

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Harcourt Brace Jovanovich College Publishers.

Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135. https://doi.org/10.1177/0265532215582282

Davis, M. T. (2010). Assessing technical communication within engineering contexts tutorial. *IEEE Transactions on Professional Communication, 53*(1), 33-45. https://doi.org/10.1109/tpc.2009.2038736

Drubin, D. G., & Kellogg, D. R. (2012). English as the universal language of science: Opportunities and challenges. *Molecular Biology of the Cell, 23*(8), 1399. https://doi.org/10.1091/mbc.e12-02-0108

Ekmekçi, E. (2016). Comparison of native and non-native English language teachers' evaluation of EFL learners' speaking skills: Conflicting or identical rating behaviour?. *English Language Teaching, 9*(5), 98-105. https://doi.org/10.5539/elt.v9n5p98

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology, 11*, Article 330. 1-14. https://doi.org/10.3389/fpsyg.2020.00330

Gan, Z. (2013). Understanding English speaking difficulties: An investigation of two Chinese populations. *Journal of Multilingual and*

Multicultural Development, 34(3), 231-248. https://doi.org/10.1080/01434632.2013.768622

Hidri, S. (2018). Assessing spoken language ability: A many-facet Rasch analysis. In S. Hidri (Ed.), *Revisiting the assessment of second language abilities: From theory to practice* (pp. 29-53). Springer, Cham. https://doi.org/10.1007/978-3-319-62884-4_2

Huang, L., Kubelec, S., Keng, N., & Hsu, L. (2018). Evaluating CEFR rater performance through the analysis of spoken learner corpora. *Language Testing in Asia, 8*(14). 1-17. https://doi.org/10.1186/s40468-018-0069-0

Iberri-Shea, G., & Hui, S. K. F. (2017). Adaptation and assessment of a public speaking rating scale. *Cogent Education, 4*(1), 1-16. https://doi.org/10.1080/2331186x.2017.1287390

Jason, F., & Xun, Y. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology, 11,* 1-14. https://doi.org/10.3389/fpsyg.2020.00330

Kaewpet, C., & Sukamolson, S. (2011). A sociolinguistic approach to oral and written communication for engineering students. *Asian Social Science, 7*(10), 183-187. https://doi.org/10.5539/ass.v7n10p183

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Lamprianou, I., Tsagari, D., & Kyriakou, N. (2021). The longitudinal stability of rating characteristics in an EFL examination: Methodological and substantive considerations. *Language Testing, 38*(2), 273-301. https://doi.org/10.1177/0265532220940960

Lee, Y. J. (2007). The multimedia assisted test of English speaking: The SOPI approach. *Language Assessment Quarterly, 4*(4), 352-366. https://doi.org/10.1080/15434300701533661

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246-276. https://doi.org/10.1191/0265532202lt230oa

Leung, L. (2015). Validity, reliability, and generalizability in qualitative research. *Journal of Family Medicine and Primary Care, 4*(3), 324-327. https://doi.org/10.4103/2249-4863.161306

Naphon, K. (2017). Presentation assessment rubric development and inter-rater reliability of communication and presentation skills course. *Journal of Humanities and Social Sciences, 9*(18), 1–18. https://ejournals.swu.ac.th/index.php/swurd/article/view/9571

Nunnally, J. D. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Orr, T. (2010). Assessment in professional communication. *IEEE Transactions on Professional Communication, 53*(1), 1-3. https://ieeexplore.ieee.org/document/5419148/

Putri, N. S. E., Pratolo, B. W., & Setiani, F. (2019). The alternative assessment of EFL students' oral competence: Practices and constraints. *Ethical Lingua: Journal of Language Teaching and Literature, 6*(2), 72-85. https://doi.org/10.30605/25409190.v6.72-85

Rubin, R. B., Welch, S. A., & Buerkel, R. A. (1995). Performance-based assessment of high school speech instruction. *Communication Education, 44*(1), 30-39. https://doi.org/10.1080/03634529509378995

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Sage Publications.

Stolarova, M., Wolf, C., Rinker, T., & Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement and correlation of ratings: An exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in psychology, 5*, 1-13. https://doi.org/10.3389/fpsyg.2014.00509

Sundqvist, P., Sandlund, E., Skar, G. B., & Tengberg, M. (2020). Effects of rater training on the assessment of L2 English oral proficiency. *Nordic Journal of Modern Language Methodology, 8*(1), 3-29. https://doi.org/10.46364/njmlm.v8i1.605

Tran, Y., & Hang, T. T. M. (2021). Use of posters to promote speaking performance among non-English majors at Thai Nguyen University of Education, Vietnam. *International Journal of Language and Literary Studies, 3*(2), 81-96. https://doi.org/10.36892/ijlls.v3i2.585

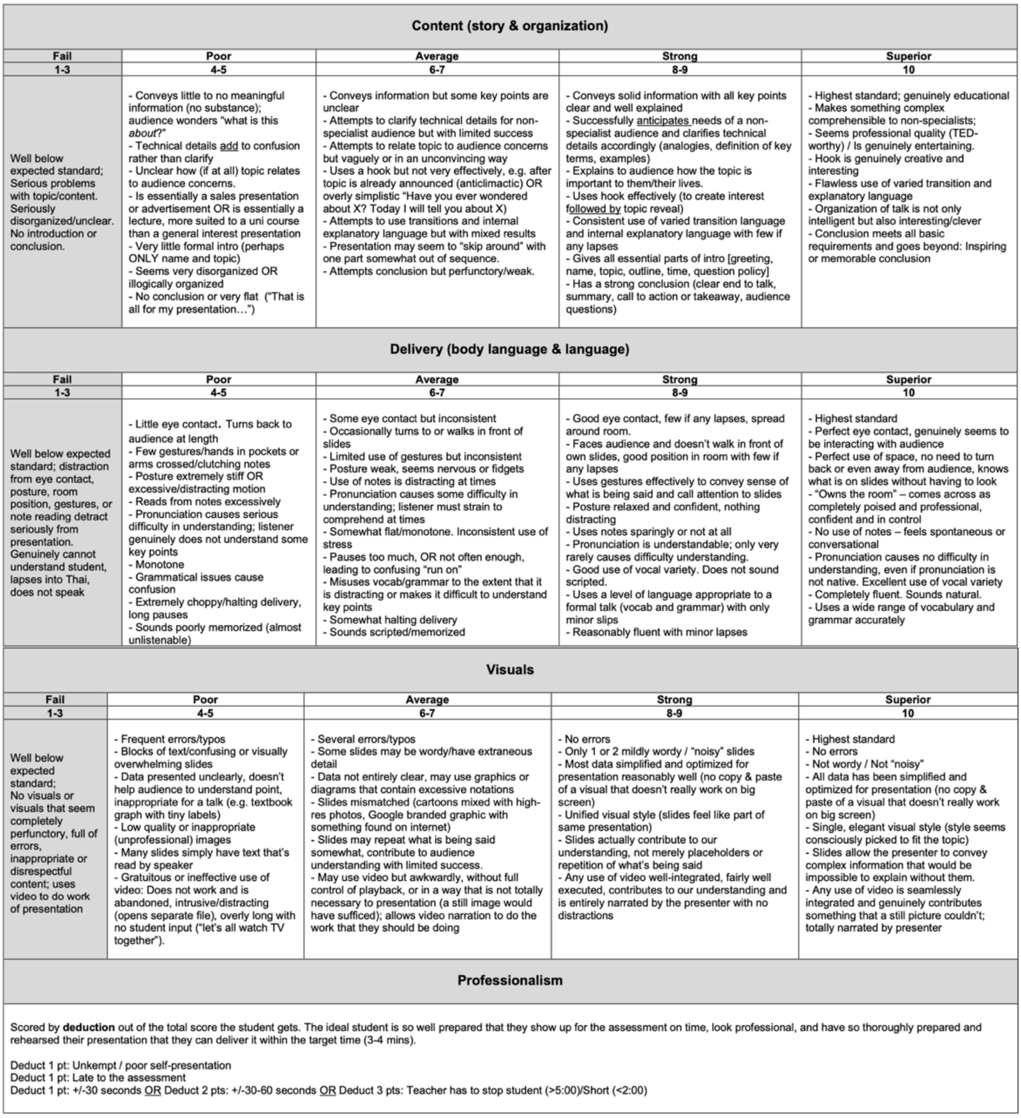
Ugiljon, A. (2018). The effective speaking testing techniques in teaching English. *International Journal of Secondary Education, 6*(1), 24-28. https://doi.org/10.11648/j.ijsedu.20180601.15

Vacha-Haase, T., (1998). Reliability generalization: Exploring variance in measurement error affecting score reliability across studies. *Educational and Psychological Measurement, 58*(1), 6-20. https://doi.org/10.1177/0013164498058001002

Wind, S. A., & Peterson, M. E. (2017). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing, 35*(2), 161-192. https://doi.org/10.1177/0265532216686999

Appendix

Solo Presentation Rubric



Content (story & organization)

Fail	Poor	Average	Strong	Superior
1-3	4-5	6-7	8-9	10
Well below expected standard; Serious problems with topic/content. Seriously disorganized/unclear. No introduction or conclusion.	- Conveys little to no meaningful information (no substance); audience wonders "what is this *about*?" - Technical details add to confusion rather than clarify - Unclear how (if at all) topic relates to audience concerns. - Is essentially a sales presentation or advertisement OR is essentially a lecture, more suited to a uni course than a general interest presentation - Very little formal intro (perhaps ONLY name and topic) - Seems very disorganized OR illogically organized - No conclusion or very flat ("That is all for my presentation…")	- Conveys information but some key points are unclear - Attempts to clarify technical details for non-specialist audience but with limited success - Attempts to relate topic to audience concerns but vaguely or in an unconvincing way - Uses a hook but not very effectively, e.g. after topic is already announced (anticlimactic) OR overly simplistic "Have you ever wondered about X? Today I will tell you about X) - Attempts to use transitions and internal explanatory language but with mixed results - Presentation may seem to "skip around" with one part somewhat out of sequence. - Attempts conclusion but perfunctory/weak.	- Conveys solid information with all key points clear and well explained - Successfully anticipates needs of a non-specialist audience and clarifies technical details accordingly (analogies, definition of key terms, examples) - Explains to audience how the topic is important to them/their lives. - Uses hook effectively (to create interest followed by topic reveal) - Consistent use of varied transition language and internal explanatory language with few if any lapses - Gives all essential parts of intro [greeting, name, topic, outline, time, question policy] - Has a strong conclusion (clear end to talk, summary, call to action or takeaway, audience questions)	- Highest standard; genuinely educational - Makes something complex comprehensible to non-specialists; - Seems professional quality (TED-worthy) / Is genuinely entertaining. - Hook is genuinely creative and interesting - Flawless use of varied transition and explanatory language - Organization of talk is not only intelligent but also interesting/clever - Conclusion meets all basic requirements and goes beyond: Inspiring or memorable conclusion

Delivery (body language & language)

Fail	Poor	Average	Strong	Superior
1-3	4-5	6-7	8-9	10
Well below expected standard; distraction from eye contact, posture, room position, gestures, or note reading detract seriously from presentation. Genuinely cannot understand student, lapses into Thai, does not speak	- Little eye contact. Turns back to audience at length - Few gestures/hands in pockets or arms crossed/clutching notes - Posture extremely stiff OR excessive/distracting motion - Reads from notes excessively - Pronunciation causes serious difficulty in understanding; listener genuinely does not understand some key points - Monotone - Grammatical issues cause confusion - Extremely choppy/halting delivery, long pauses - Sounds poorly memorized (almost unlistenable)	- Some eye contact but inconsistent - Occasionally turns to or walks in front of slides - Limited use of gestures but inconsistent - Posture weak, seems nervous or fidgets - Use of notes is distracting at times - Pronunciation causes some difficulty in understanding; listener must strain to comprehend at times - Somewhat flat/monotone. Inconsistent use of stress - Pauses too much, OR not often enough, leading to confusing "run on" - Misuses vocab/grammar to the extent that it is distracting or makes it difficult to understand key points - Somewhat halting delivery - Sounds scripted/memorized	- Good eye contact, few if any lapses, spread around room. - Faces audience and doesn't walk in front of own slides, good position in room with few if any lapses - Uses gestures effectively to convey sense of what is being said and call attention to slides - Posture relaxed and confident, nothing distracting - Uses notes sparingly or not at all - Pronunciation is understandable; only very rarely causes difficulty understanding. - Good use of vocal variety. Does not sound scripted. - Uses a level of language appropriate to a formal talk (vocab and grammar) with only minor slips - Reasonably fluent with minor lapses	- Highest standard - Perfect eye contact, genuinely seems to be interacting with audience - Perfect use of space, no need to turn back or even away from audience, knows what is on slides without having to look - "Owns the room" – comes across as completely poised and professional, confident and in control - No use of notes – feels spontaneous or conversational - Pronunciation causes no difficulty in understanding, even if pronunciation is not native. Excellent use of vocal variety - Completely fluent. Sounds natural. - Uses a wide range of vocabulary and grammar accurately

Visuals

Fail	Poor	Average	Strong	Superior
1-3	4-5	6-7	8-9	10
Well below expected standard; No visuals or visuals that seem completely perfunctory, full of errors, inappropriate or disrespectful content; uses video to do work of presentation	- Frequent errors/typos - Blocks of text/confusing or visually overwhelming slides - Data presented unclearly, doesn't help audience to understand point, inappropriate for a talk (e.g. textbook graph with tiny labels) - Low quality or inappropriate (unprofessional) images - Many slides simply have text that's read by speaker - Gratuitous or ineffective use of video: Does not work and is abandoned, intrusive/distracting (opens separate file), overly long with no student input ("let's all watch TV together").	- Several errors/typos - Some slides may be wordy/have extraneous detail - Data not entirely clear, may use graphics or diagrams that contain excessive notations - Slides mismatched (cartoons mixed with high-res photos, Google branded graphic with something found on internet) - Slides may repeat what is being said somewhat, contribute to audience understanding with limited success. - May use video but awkwardly, without full control of playback, or in a way that is not totally necessary to presentation (a still image would have sufficed); allows video narration to do the work that they should be doing	- No errors - Only 1 or 2 mildly wordy / "noisy" slides - Most data simplified and optimized for presentation reasonably well (no copy & paste of a visual that doesn't really work on big screen) - Unified visual style (slides feel like part of same presentation) - Slides actually contribute to our understanding, not merely placeholders or repetition of what's being said - Any use of video well-integrated, fairly well executed, contributes to our understanding and is entirely narrated by the presenter with no distractions	- Highest standard - No errors - Not wordy / Not "noisy" - All data has been simplified and optimized for presentation (no copy & paste of a visual that doesn't really work on big screen) - Single, elegant visual style (style seems consciously picked to fit the topic) - Slides allow the presenter to convey complex information that would be impossible to explain without them. - Any use of video is seamlessly integrated and genuinely contributes something that a still picture couldn't; totally narrated by presenter

Professionalism

Scored by **deduction** out of the total score the student gets. The ideal student is so well prepared that they show up for the assessment on time, look professional, and have so thoroughly prepared and rehearsed their presentation that they can deliver it within the target time (3-4 mins).

Deduct 1 pt: Unkempt / poor self-presentation
Deduct 1 pt: Late to the assessment
Deduct 1 pt: +/-30 seconds OR Deduct 2 pts: +/-30-60 seconds OR Deduct 3 pts: Teacher has to stop student (>5:00)/Short (<2:00)