*PASAA Journal*
Volume 67, July – December 2023, 213–246

# Measuring EFL Students' Oral Proficiency Improvement Using Teachers' Judgment: The Effects of Two Types of Classroom Instruction

Dony Marzuki

Politeknik Negeri Padang, Padang, Indonesia
*Corresponding author: donymarzuki03@gmail.com*

| **Article information** | |
|---|---|
| **Abstract** | This study investigated the effects of two types of classroom instructions on EFL learners' oral proficiency by implementing teachers' judgment. Two intact classes of EFL university students studied a compulsory subject of their department under two different types of classroom instruction. The first group of students was trained with explicit oral strategy training instruction, which taught them to learn and practice strategies to improve their speech fluency. The second group was instructed to practice the task twice as much as the first group using implicit task-based instruction. The audio recording of both groups' oral task performance in the pre-test and post-test conditions were rated for proficiency, pronunciation, discourse, vocabulary, grammar, and complexity. Two English teachers were trained to rate the recordings based on two oral proficiency rubrics. The non-parametric tests and estimation analysis results showed that both groups' oral proficiency improved significantly, with effect sizes ranging from medium to large. The comparison of both groups' results showed that the two types of instruction had a comparable effect on students' oral proficiency. The current study's findings suggest that the use of teachers' judgments can be necessary for classroom-based |

E-ISSN: 2287-0024

| | |
|---|---|
| | studies to measure the effects of instruction on gains in learners' oral proficiency. |
| **Keywords** | explicit oral strategy training instruction, implicit task-based instruction, oral proficiency, speech fluency, teacher's judgment |
| **APA citation:** | Marzuki, D. (2023). Measuring EFL students' oral proficiency improvement using teachers' judgment: The effects of two classroom instructions. *PASAA, 67*, 213–246. |

E-ISSN: 2287-0024

## 1. Introduction

In the EFL classroom context, instruction for teaching speaking used to be transmitted from one generation of teachers to the next with little to no modifications over time. This practice occurred because the new teachers believed in the effectiveness of the instruction they adopted in facilitating and enabling learning, especially in the classroom (De Graaff & Housen, 2009). Therefore, they did not attempt to investigate further or validate the instruction. However, second and foreign language researchers are not "on the same page" regarding the important role of instruction in language learning. Some researchers believe that language learning is incidental and does not require intervention like instruction (e.g., Krashen, 1985; Long & Robinson, 1998; Reynolds et al., 2022; Thomas, 2020); others affirm that instruction could be a key element in second and foreign language learning (e.g., Ahmadian & Tavakoli, 2011; Ellis, 2005; Kang et al., 2019; Tavakoli et al., 2016; Ziegler & González-Lloret, 2022). Nevertheless, it is generally concluded that instruction has a substantial effect on learning outcomes (Norris & Ortega, 2000).

In the area of speaking and oral proficiency, the role of instruction has been recognized (Plonsky, 2011). Previous studies have reported that explicit instruction could improve learners' speaking (i.e., Sato, 2020; Seifoori & Vahidi, 2012; Tavakoli et al., 2016), while others have found that learners' oral proficiency can also be improved through implicit instruction, particularly when timing and tasks are planned properly (i.e., Bygate, 2001; Foster & Skehan, 1996; Lambert et al., 2017; Lambert et al., 2020; Wang, 2014). Considering the benefits offered by the two types of instruction—explicit and implicit—in improving EFL learners' speaking proficiency, it would be beneficial for EFL teachers to know the kind of instruction that would be more beneficial for their learners.

Research investigating the effect of instruction on oral proficiency usually applies specific criteria or measures targeting certain elements of learners' performance, such as speech accuracy, complexity, and fluency. No previous study

has used teachers' judgment to measure an instructional effect, however. To date, teachers' judgment is only applied to research to score learners' oral performance, not to determine the effects of instruction, based on specific criteria or scales. Those research studies usually focus on investigating the variability among raters (Isaacs & Thomson, 2013; Mohd Noh & Mohd Matore, 2022; Winke et al., 2013), including the influence of raters' accent on ratings (e.g., Carey et al., 2011; Derwing & Munro, 1997; Huang et al., 2016; Xu et al., 2023). Therefore, the findings of the current study offer an alternative to measure the effectiveness of classroom-based instruction by applying a teacher's judgment approach. The results of this study have theoretical and practical implications in EFL instruction for the teaching of speaking.

## 2. Literature Review

Ellis et al. (2009) suggest that implicit instruction should be able to stimulate students to infer language rules without asking them to pay attention to the rules. Implicit instruction presents the target features in a contextualized manner by simply presenting the target information or the language features to students. They would then be encouraged to conclude or create their own conceptual structures by using their linguistic repertoires. In implicit instruction, teachers function as learning facilitators rather than instructors. On the contrary, explicit instruction is a teaching approach that draws students' attention to a target feature and teaches students a specific language rule (DeKeyser, 1995). In explicit grammar teaching, for example, teachers provide an explanation about the grammatical rules being learned. In implicit instruction, the students are encouraged to find the grammatical rules by themselves. This process could be eased when teachers provide comprehensive language input and practice or select pedagogic tasks that will accommodate students' learning. It is worth noting that Chun et al. (2011) conclude that the central construct associated with implicit and explicit instruction is students' attention. Implicit instruction should not direct students' attention during learning, while explicit instruction should do otherwise.

A body of research has been devoted to investigating the effects of implicit task-based instruction (i.e., Bygate, 2018; Hsu, 2019; Ahmadian et al., 2015) on EFL learners' oral proficiency (e.g., Bosker et al., 2013; Bui et al., 2019; Lambert et al., 2017; Lambert et al., 2020). For example, to reveal the benefit of implicit instruction, Lambert et al. (2020) used four pre-task planning conditions to improve Japanese EFL students' oral performance. The study found that all four planning conditions helped students in the speech production process by easing the demand of conceptualization, formulation, and monitoring, which, in the end, enabled the students to produce more fluent speech. In another study, de Jong and Perfetti (2011) investigated the effects of task repetition on two groups of L2 learners. The first group was assigned to perform one task three times, while the others had to perform three tasks without repetition. The results showed that task repetition had a noticeable effect on students' speech fluency.

Also dealing with task repetition, Lambert et al. (2017) investigated the effect of the repetition of aural-oral monologue tasks on immediate gains in L2 fluency of Japanese university students in a classroom setting. The students were assigned to complete three oral tasks six times. Results revealed that task repetition was related to gains in oral fluency regardless of proficiency level or task type. Meanwhile, Bui et al. (2019) used five task repetition conditions in their study and investigated the effects on EFL learners' oral performance. The participants were divided according to the five different task repetition conditions: immediate, one-day, three-day, one-week, and two-week time-space repetitions. The study found that all conditions of task repetition had a positive effect on oral performance, with speed fluency benefitting the most from immediate and small intervals of time-space between initial and repeated performance.

Explicit instruction has been applied to improve learners' listening (i.e., Dalman & Plonsky, 2022; Fathi & Hamidizadeh, 2019; Milliner & Dimoski, 2021), reading (i.e., Brevik, 2019; Fathi & Afzali, 2020; Wang, 2016), writing (i.e., Rietdijk et al., 2018; Teng & Huang, 2019; Teng & Zhang, 2020), and speaking (i.e., Chou,

2018; Dao, 2020; Forbes & Fisher, 2018; Sato, 2020). Some studies have used strategy training instruction as their method (e.g., Brevik, 2019; Dao, 2020; Nakatani, 2005; Sato, 2020; Suzuki, 2021; Tavakoli et al., 2016; Teng & Zhang, 2020). Most of these studies have found that explicit instruction can have an effect on L2 and EFL learners' language skills improvement. Furthermore, Sato (2020) found the effects of socio-affective strategy training on L2 oral performance when investigating the effect of explicit metacognitive instruction for collaborative interaction (MICI) during communicative tasks. He also compared metacognitive instruction with implicit task-based instruction. The study revealed that learners who received strategy-based instruction outperformed learners who received implicit task-based instruction in strategy use and comprehensibility of their discourse.

In addition, Nakatani (2005) investigated the effects of oral communication strategies (socio-affective) on learners' oral proficiency. During a 12-week course, learners were trained with metacognitive strategies focused on oral communication strategies such as help-seeking, time-gaining, and negotiation of meaning. It was reported that the training successfully improved learners' oral proficiency with the dominant use of specific strategies such as maintaining fluency and negotiation of meaning. Finally, a short intervention study carried out by Tavakoli et al. (2016) to examine the effects of pedagogic intervention in fluency strategy training on developing fluency among L2 learners showed that after the four meetings, which were conducted within a four-week duration, the participants improved their speech fluency.

It is worth noting that most of the previous implicit and explicit instruction studies were conducted to investigate the effects of instruction on improving L2 and EFL learners' language proficiency, performance, and acquisition. In oral proficiency, these studies usually applied specific criteria or measures targeting some aspects of learners' performance such as speech accuracy, complexity, and fluency. The use of teachers' judgment to measure instructional effects has never

been applied. Judging from their expertise regarding their teaching topics and familiarity with assessing and grading learners' assignments, teachers should be able to measure the effects of applied classroom instruction. The current study aimed at investigating the effects of two types of classroom instruction on EFL learners' oral proficiency by implementing teachers' judgment. Specifically, two experienced English teachers were employed as raters to judge oral proficiency development of two groups of EFL students who received different types of classroom instruction.

## 3. Methodology

This study investigated the effects of two types of classroom instruction, which were explicit strategy training instruction and implicit task-based instruction, on EFL students' oral proficiency. The three research questions addressed in this study were as follows:

1. Does explicit strategy training instruction affect EFL students' oral proficiency?
2. Does implicit task-based instruction affect EFL students' oral proficiency?
3. Do explicit and implicit instruction have different effects on EFL students' oral proficiency?

### 3.1 Research Design

The present study was quasi-experimental research with a two-group pre-test and post-test design.

### 3.2 Participants

The participants in this study were 54 EFL students from two intact classes of an English Department at an Indonesian university. Participants' English proficiency as measured by the Test of English as International Communication (TOEIC) ranged from 240 to 655, which is approximately equivalent to a Common European Framework of Reference for Languages (CEFR) level from Pre-intermediate or Basic User (A2) to Intermediate or Independent User (B1) ( Council

of Europe, 2001). The two classes served as the two groups of the study, with 29 students in the explicit strategy training instruction group (EG) and 25 students in the implicit task-based instruction group (IG). Both groups were comparable in terms of students' level of proficiency, age, and gender distribution.

### 3.3 Materials

Three forms of pedagogic tasks were used as the study materials. The main task was an output-based monologic news report that was meant to elicit participants' speech samples. This task required participants to work with a YouTube news video and was labeled as Task 3. Meanwhile, Task 1 and Task 2 were two preliminary tasks intended to prepare the participants for Task 3. Task 1 was input-based and called *identifying factual errors in a news report*. Task 2 was output-based and named *retelling a news report to a friend.*

In total, the participants performed eight monologic news report tasks during the study's eight instructional meetings. The news reports were based on eight YouTube news videos which dealt with (1) a terrorist incident, (2) a road accident, (3) a natural disaster, (4) an airplane accident, (5) a forest fire, (6) an accident at sea, (7) domestic violence, and (8) a hate crime. The pre-test and the post-test also used Task 3 with different news videos. The videos were a terrorist attack in Indonesia for the pre-test and a natural disaster in Indonesia for the post-test.

It is noteworthy that in this study, raters' judgments were used to provide an independent measure of participants' oral proficiency gains. Two experienced English teachers judged participants' performances in the pre-test and post-test based on Brown's (2001) and Choi's (2005) oral proficiency rating scales. Brown's holistic rating covers only the general proficiency scale, while Choi's analytic rating includes five independent scales: pronunciation, discourse, vocabulary, grammar, and complexity. The results from both raters were then compared to ascertain if there was any meaningful

improvement in participants' oral performance when comparing their pre-test to their post-test scores.

### 3.4 Procedures

Eight instructional meetings with different news topics were provided for both groups. Each meeting consisted of three learning stages, which lasted two hours. The stages were input-based, task preparation, and output-based. The instruction was started with the input-based stage. This stage was intended to prepare the participants for the lesson, especially the topic being discussed. They watched a video twice and were instructed to get as much information as they could from it. After that, they worked on a fact sheet to identify and fix any incorrect statements from the video.

In the second stage, the task preparation stage, the two groups went through different treatments. The EG started their fluency strategy training, while the IG began their task practice. The training was focused on utilizing aspects of utterance fluency (speed, breakdown, and repair) which were incorporated into specific strategies such as (1) paying attention to personal patterns of frequent pausing, (2) practical use of single-word and lexical-chunk fillers, (3) avoiding repetition, (4) avoiding false starts, (5) avoiding reformulation, and (6) avoiding replacement. These strategies were then distributed across the eight instructional meetings. After the training, the EG continued to work on Task 2.

The only difference between the EG and the IG during this stage was the number of tasks each group performed. The IG performed Task 2 twice with two different topics, while the EG only did it once.

Meanwhile, in the output-based stage, the lesson was the same for both groups. Here, the participants performed Task 3. They were asked to perform a 1-minute news report speech based on a YouTube news video. They performed the task three times by recording it on a computer. They were also encouraged to think about it and conduct the task carefully as if their report might be broadcasted on the university's radio station. This stage began with the participants watching the

news video twice, during which they could make notes. They could use the notes for their first two recordings but not for their third one. Therefore, all notes were taken by the teacher before they recorded their speech for the third time. This final recording was used for the analysis

### 3.5 Data Analysis

Audio recordings of the participants' pre-test and post-test were collected and transcribed for analysis purposes. The transcription was used as an additional reference for the two raters to judge participants' performance. The raters were two English teachers at the institution. They had more than five-years of experience teaching speaking courses to students. To maintain the judgment objectivity, all participants were anonymous.

The non-parametric tests of the inferential analysis were employed in this study because the data were found to be highly skewed and not normally distributed. The common data transformational tools, such as logarithmic and square root, failed to improve the data. Mann-Whitney U and Wilcoxon Signed Rank tests were then used. In addition, estimation analysis was used to explain the magnitude and precision of an effect being investigated. Estimation analysis focuses on the estimation of effect sizes or the point of estimates and their confidence intervals (precision estimates) (Claridge-Chang & Assam, 2016; Cumming, 2014). This analysis method was considered useful in explaining the results of the present study that revealed non-significant effects as information regarding the magnitude and precision of the effects would be considered meaningful as well (Claridge-Chang & Assam, 2016).

## 4. Results

### 4.1 Raters' Judgment for Explicit Group (EG)

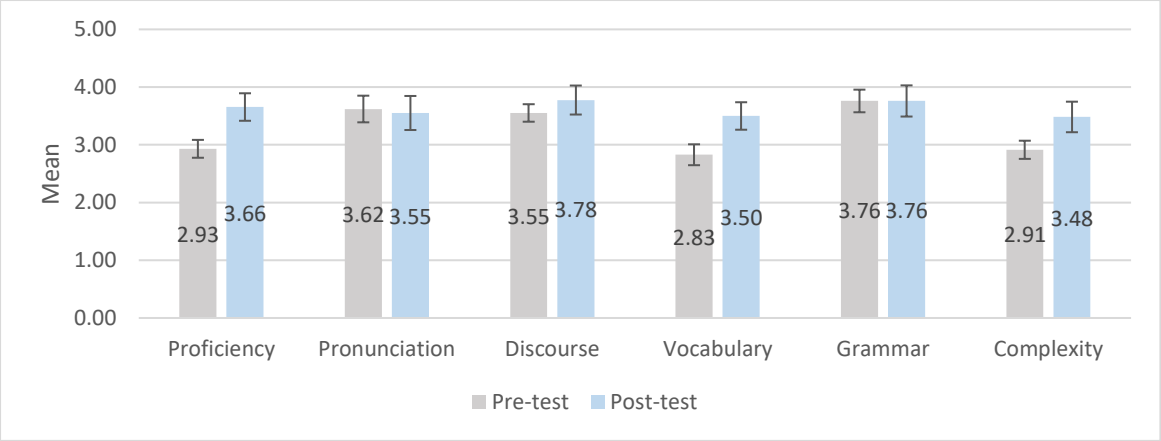
For analysis purposes, the scales from both raters were combined and divided by two (the number of raters) to obtain the average mean scores. These mean scores were then analyzed using a bar chart for visual analysis, the Wilcoxon

Signed Rank Test, for the significance test, and the paired mean difference estimation of ESCI to determine meaningful differences between the pre-test and the post-test scores of the EG.

First, a bar chart accompanied by the error bars representing the 95% confidence intervals of the scales was created (Figure 1).

**Figure 1**

*The Explicit Group's Pre-Test and Post-Test Mean Scores for the Oral Proficiency Scales*




As can be seen in Figure 1, for most of the scales, there was a visible increase in the mean from the pre-test to the post-test, except for pronunciation. The general proficiency scale appeared to have the biggest increase, followed by vocabulary and complexity. To investigate whether the visible increases in the mean scores of the oral proficiency scales were statistically significant, the Wilcoxon Signed Rank Test was conducted, and the results are presented in Table 1.

**Table 1**

*The results of the Wilcoxon Signed Rank Test of the Explicit Group's Oral Proficiency Scales*

| Variables | Tests | N | Mean Rank | *z* | *p* |
|---|---|---|---|---|---|
| Proficiency | Pre-test | 29 | 1.14 | -3.73 | < .01 |
| | Post-test | 28 | 1.86 | | |
| | Total | 57 | | | |
| Pronunciation | Pre-test | 29 | 1.52 | -0.46 | .65 |
| | Post-test | 28 | 1.48 | | |
| | Total | 57 | | | |
| Discourse | Pre-test | 29 | 1.29* | -2.08 | .04 |
| | Post-test | 28 | 1.71* | | |
| | Total | 57 | | | |
| Vocabulary | Pre-test | 29 | 1.12* | -3.77 | < .01 |
| | Post-test | 28 | 1.88* | | |
| | Total | 57 | | | |
| Grammar | Pre-test | 29 | 1.41 | -0.49 | .62 |
| | Post-test | 28 | 1.59 | | |
| | Total | 57 | | | |
| Complexity | Pre-test | 29 | 1.10* | -3.20 | < .01 |
| | Post-test | 28 | 1.90* | | |
| | Total | 57 | | | |

*Note: The mean difference was significant at the .05 level.*

It can be seen from Table 1 that the four scales in the list resulted in a statistically significant difference. These scales were proficiency ($z$ = -3.73, $p$ < .01), discourse ($z$ = -2.08, $p$ = .04), vocabulary ($z$ = -3.77, $p$ < 0.1), and complexity ($z$ = -3.20, $p$ < .01). The average observed power of these scales was acceptable (> .80), except for discourse, whose observed power was only .55. The differences in two other scales, pronunciation and grammar, did not reach statistical significance. These results support the visual analysis results of the bar chart (Figure 1), indicating that the EG's oral proficiency improvement was the effect of the strategy training instruction.

Following the test of significance, a paired mean difference estimation analysis was conducted on the general proficiency scales to determine the magnitudes of the effects (effect sizes). The estimation analysis is detailed in Table 2.

**Table 2**

*The Results of Paired Mean Difference Estimation for Oral Fluency Scales for the Explicit Group*

| Variables | Condition | M | 95 % CI | | SD | $d_{average}$ | N |
|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | | | |
| Proficiency | Post-test | 3.66 | 3.41 | 3.90 | 0.64 | | 29 |
| | Pre-test | 2.93 | 2.77 | 3.09 | 0.42 | | 29 |
| | Difference | 0.72 | 0.46 | 0.99 | 0.69 | 1.32 | 29 |
| Pronunciation | Post-test | 3.55 | 3.25 | 3.85 | 0.79 | | 29 |
| | Pre-test | 3.62 | 3.38 | 3.86 | 0.62 | | 29 |
| | Difference | -0.07 | -0.32 | 0.18 | 0.66 | 0.10 | 29 |
| Discourse | Post-test | 3.78 | 3.52 | 4.03 | 0.68 | | 29 |
| | Pre-test | 3.55 | 3.40 | 3.71 | 0.41 | | 29 |
| | Difference | 0.22 | 0.01 | 0.44 | 0.56 | 0.40 | 29 |
| Vocabulary | Post-test | 3.50 | 3.26 | 3.74 | 0.64 | | 29 |
| | Pre-test | 2.83 | 2.64 | 3.01 | 0.49 | | 29 |
| | Difference | 0.67 | 0.43 | 0.92 | 0.65 | 1.17 | 29 |
| Grammar | Post-test | 3.76 | 3.48 | 4.04 | 0.73 | | 29 |
| | Pre-test | 3.76 | 3.56 | 3.96 | 0.53 | | 29 |
| | Difference | 0.00 | -0.27 | 0.27 | 0.71 | 0 | 29 |
| Complexity | Post-test | 3.48 | 3.21 | 3.75 | 0.71 | | 29 |
| | Pre-test | 2.91 | 2.75 | 3.08 | 0.42 | | 29 |
| | Difference | 0.57 | 0.28 | 0.86 | 0.75 | 0.96 | 29 |

*Note: The standardized effect size was $d_{average}$ because the denominator used was SDavg. The standardized effect size was corrected for bias. The bias-corrected version of Cohen's d was also referred to as Hedges' g. Minus values in $d_{average}$ were caused by higher scores in the pre-test than in the post-test.*

It can be seen from Table 2 that the four scales in the list had a meaningful difference between the pre-test and the post-test. The effect sizes average of the scales ranged from .4 (near medium, for discourse) to 1.32 (large, for proficiency). The general proficiency scale appeared to have the biggest mean difference ($M_{dif}$ = .72 [.48, .99]) and effect size ($d_{avg}$ = 1.32). In the pre-test, participants' mean score on the general proficiency scale was $M$ = 2.93. This score was within Level 2 of the rubric, and the speaking skill at this level is called developing speaking (Brown, 2001). In the post-test, the mean of general proficiency mean improved to Level 3 ($M$ = 3.66), which is categorized as competent speaking. The second scale, representing a very large effect size ($d_{avg}$ = 1.17), was vocabulary ($M_{dif}$ = .67 [.43, .92]). In this aspect, participants' speaking skills improved from Level 2 ($M$ = 2.83) in the pre-test to Level 3 ($M$ = 3.50). Level 2 in this analytic rubric meant that participants' vocabulary was less varied, with many words used repeatedly. Level 3 indicated varied vocabulary with some use of idiomatic expressions. The other scale, which also resulted in a large effect size, was complexity ($d_{avg}$ = .96, $M_{dif}$ = .57 [.28, .86]). The speaking skill in this scale similarly improved from Level 2 ($M$ = 2.91) to Level 3 ($M$ = 3.48).

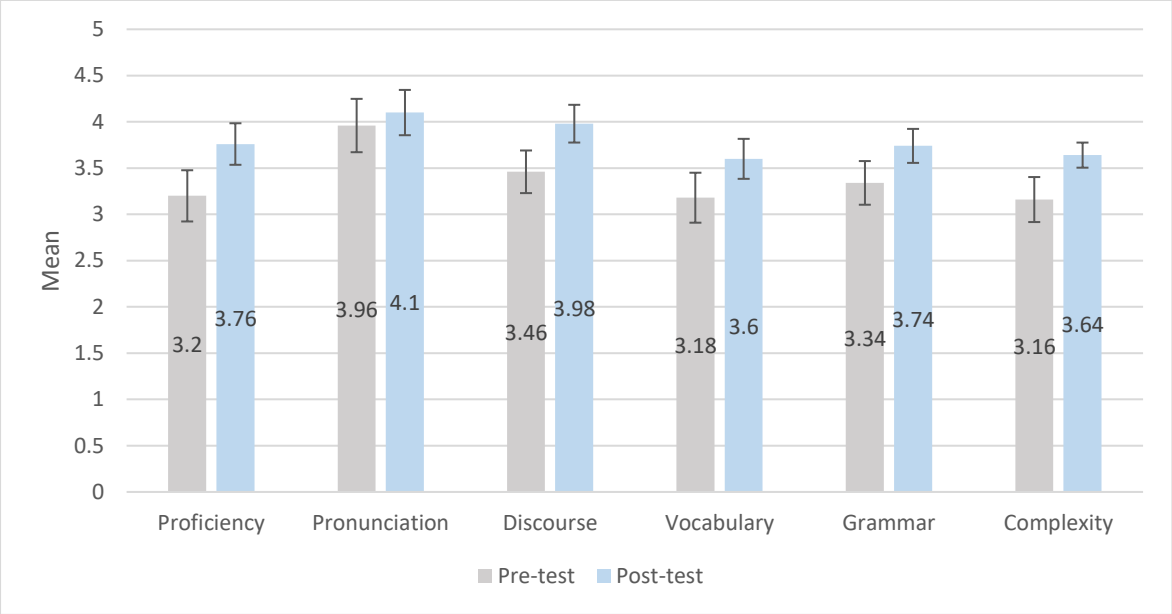
A different trend was shown for the other two scales, pronunciation and grammar, since their scores did not increase between the pre-test and the post-test. In fact, pronunciation scores decreased. Apart from these two scores, it can be concluded that based on the raters' judgments, the participants from this group improved somewhat in their oral proficiency.

### 4.2 Raters' Judgment for Implicit Group (IG)

The same two raters also judged participants in the IG based on the same oral proficiency rating scales. Similar to the analysis performed for the EG, the analysis of oral fluency scales for the IG also involved bar charts, a test of significant difference, and estimation analysis. The bar chart was used to visualize the difference between the pre-test and post-test means for the IG and is presented in Figure 2.

**Figure 2**

*The Implicit Group's Pre-Test and Post-Test Mean Scores for the Oral Proficiency Scales*




For most of the oral proficiency scales in Figure 2, an increase in the mean from the pre-test to the post-test was visible, with general proficiency and discourse seeming to improve the most. A visible increase was also evident for complexity, while in the other three scales (pronunciation, vocabulary, and grammar), the increase seemed moderate. Next, the Wilcoxon Signed Rank Test was conducted to investigate the significance of this mean difference (increase) between the pre-test and post-test scores for this IG (see Table 3).

**Table 3**

*The Results of the Wilcoxon Signed Rank Test for the Implicit Group's Oral Proficiency Scales*

| Variables | Tests | N | Mean Rank | *z* | *Sig.* (*p*) |
|---|---|---|---|---|---|
| Proficiency | Pre-test | 25 | 1.16* | -3.67 | < .01 |
| | Post-test | 25 | 1.84* | | |
| | Total | 50 | | | |
| Pronunciation | Pre-test | 25 | 1.48 | -1.07 | .28 |
| | Post-test | 25 | 1.52 | | |
| | Total | 50 | | | |
| Discourse | Pre-test | 25 | 1.20* | -3.49 | < .01 |
| | Post-test | 25 | 1.80* | | |
| | Total | 50 | | | |
| Vocabulary | Pre-test | 25 | 1.26* | -3.08 | < .01 |
| | Post-test | 25 | 1.74* | | |
| | Total | 50 | | | |
| Grammar | Pre-test | 25 | 1.22* | -3.35 | < .01 |
| | Post-test | 25 | 1.78* | | |
| | Total | 50 | | | |
| Complexity | Pre-test | 25 | 1.20* | -3.41 | < .01 |
| | Post-test | 25 | 1.80* | | |
| | Total | 50 | | | |

*Note: * The mean difference was significant at the 0.05 level.*

Figure 2 reveals a statistically significant difference in means for five variables between the pre-test and post-test. The level of significance (*p*) of these five was less than .01. The observed power was also larger than .80. Pronunciation was the only variable for which there was not any significant difference ($z = -1.07$, $p = .28$). The results indicated that the implicit instruction improved the participants' oral proficiency.

To determine the magnitudes of the effects (effect sizes), a paired mean difference estimation analysis using ESCI was conducted on the oral fluency scales for this IG (see Table 4).

**Table 4**

*The Results of Paired Mean Difference Estimation for Oral Fluency Scales for the Implicit Group*

| Variables | Condition | M | 95 % CI | | SD | $d_{average}$ | N |
|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | | | |
| Proficiency | Post-test | 3.76 | 3.53 | 3.99 | 0.56 | | 25 |
| | Pre-test | 3.20 | 2.91 | 3.49 | 0.69 | | 25 |
| | Difference | 0.56 | 0.34 | 0.78 | 0.53 | 0.87 | 25 |
| Pronunciation | Post-test | 4.10 | 3.85 | 4.35 | 0.61 | | 25 |
| | Pre-test | 3.96 | 3.66 | 4.26 | 0.72 | | 25 |
| | Difference | 0.14 | -0.12 | 0.40 | 0.62 | 0.21 | 25 |
| Discourse | Post-test | 3.98 | 3.77 | 4.19 | 0.51 | | 25 |
| | Pre-test | 3.46 | 3.22 | 3.70 | 0.58 | | 25 |
| | Difference | 0.52 | 0.30 | 0.74 | 0.53 | 0.94 | 25 |
| Vocabulary | Post-test | 3.60 | 3.38 | 3.82 | 0.54 | | 25 |
| | Pre-test | 3.18 | 2.90 | 3.46 | 0.68 | | 25 |
| | Difference | 0.42 | 0.18 | 0.66 | 0.57 | 0.68 | 25 |
| Grammar | Post-test | 3.74 | 3.55 | 3.93 | 0.46 | | 25 |
| | Pre-test | 3.34 | 3.10 | 3.58 | 0.59 | | 25 |
| | Difference | 0.40 | 0.19 | 0.61 | 0.50 | 0.74 | 25 |
| Complexity | Post-test | 3.64 | 3.50 | 3.78 | 0.34 | | 25 |
| | Pre-test | 3.16 | 2.91 | 3.41 | 0.61 | | 25 |
| | Difference | 0.48 | 0.25 | 0.71 | 0.55 | 0.96 | 25 |

*Note: The standardized effect size was $d_{average}$ because the denominator used was SDavg. The standardized effect size was corrected for bias. The bias-corrected version of Cohen's d was alsocalled Hedges' g. Minus values in $d_{average}$ was caused by higher scores in the pre-test than the post-test.*

Table 4 shows that a potentially meaningful difference between the pre-test and the post-test exists for the five scales. The effect sizes ranged from .68 (medium, for vocabulary) to .96 (large, for complexity). For participants in this IG, the complexity scale had the largest effect size ($d_{average}$ = .96) and the second biggest mean difference ($M_{dif}$ = .48 [.25, .71]). Improvement in participants'

speaking skills in this aspect was still within Level 3 ($M$ = 3.16 in the pre-test to $M$ = 3.64 in post-test).

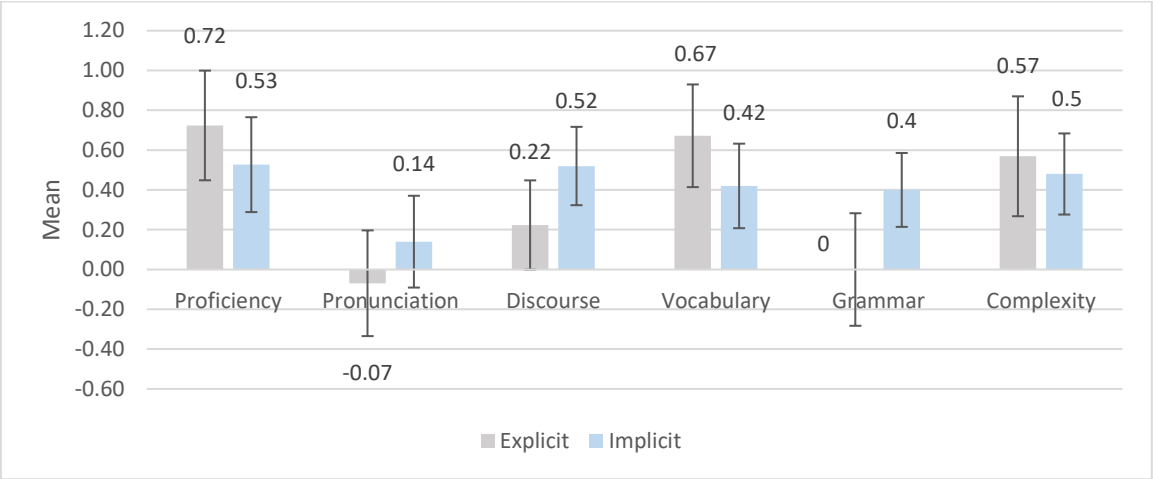
Similar to the EG, the general proficiency scale for the IG manifested the largest mean difference between the pre-test and post-test ($M_{dif}$ = .56 [.34, .78]) with a large effect size ($d_{average}$ = .87). In the pre-test, participants' mean score on the general proficiency scale was $M$ = 3.20 (Level 3). In the post-test, the general proficiency's mean improved to $M$ = 3.76 (Level 3). Meanwhile, pronunciation was the lowest among all of the scales in terms of the effect size ($d_{average}$ = .21) and mean difference ($M_{dif}$ = .14 [-.12, .40]) between the pre-test ($M$ = 3.96) and post-test ($M$ = 4.10). Therefore, it can be concluded that, based on the raters' judgment, there was a meaningful improvement in participants' oral proficiency. Although this improvement did not shift from one level to another (except for pronunciation), the mean difference between the pre-test and post-test scores for the IG was comparable to the EG. For EFL learners in the context of the current study, the extent of the mean differences and effect sizes reflected a shift in their oral proficiency development.

### 4.1. The Different Effects of Explicit and Implicit Instruction

Regarding the comparison of oral proficiency gains between groups, the same analyses were also applied. Similar to the first analysis, bar charts were used to enable a visual inspection of the gains (see Figure 3).

## Figure 3

*Both Groups' Oral Proficiency Gains*




Note: 1 denotes the explicit group, and 2 is the implicit group. Error bars represent 95% confidence intervals of the means.

Figure 3 shows that the IG had a higher gain than the EG in three scales: pronunciation, discourse, and grammar. The EG experienced a setback in their pronunciation, while there was no gain for their grammar since the mean was zero. The highest difference in gain between the two groups was found for grammar. On the other hand, the EG's gains were higher than that of the IG in terms of the other three scales: general proficiency, vocabulary, and complexity. The gain in vocabulary was the greatest difference between the groups.

Further analysis using the Mann-Whitney U Test was then performed to investigate the level of significance of the differences shown by the bar chart in Figure 3. The results of the Mann-Whitney U Test are presented in Table 5.

**Table 5**

*The Results of the Mann-Whitney U Test of Oral Proficiency Gain for Both Groups*

| Variables | Tests | N | Mean Rank | *z* | *Sig.* (*p*) |
|---|---|---|---|---|---|
| Proficiency | Explicit | 28 | 30.21 | -1.41 | .16 |
| | Implicit | 25 | 24.36 | | |
| | Total | 53 | | | |
| Pronunciation | Explicit | 28 | 26.12 | -0.72 | .47 |
| | Implicit | 25 | 29.10 | | |
| | Total | 53 | | | |
| Discourse | Explicit | 28 | 24.59 | -1.55 | .12 |
| | Implicit | 25 | 30.88 | | |
| | Total | 53 | | | |
| Vocabulary | Explicit | 28 | 31.48* | -2.08 | .04 |
| | Implicit | 25 | 22.88* | | |
| | Total | 53 | | | |
| Grammar | Explicit | 28 | 23.69* | -2.03 | .04 |
| | Implicit | 25 | 31.92* | | |
| | Total | 53 | | | |
| Complexity | Explicit | 28 | 30.10 | -1.39 | .16 |
| | Implicit | 25 | 24.48 | | |
| | Total | 53 | | | |

*Note: * The mean different was significant at the 0.05 level.*

As reported in Table 5, the different in gains between the two instructional groups reached a statistical significance in two scales: vocabulary ($z$ = -2.08, $p$ = .04) and grammar ($z$ = -2.03, $p$ = .04). However, the observed power for these two scales (.32 for vocabulary and .68 for grammar) was lower than .8, which was the acceptable value. Therefore, while these results were statistically significant, they were considered inconclusive. This may offer further support for the view that the effect of explicit strategy training instruction was comparable to the effect of implicit instruction in improving EFL students' oral proficiency.

In order to determine the magnitude of the difference between both groups' oral proficiency gains, estimation analysis was performed on the six oral proficiency scales (see Table 6).

**Table 6**

*The Results of Independent Mean Difference Analysis on Raters' Judgment for Gains: A Comparison Between the Explicit and Implicit Groups*

| Variables | Condition | M | 95 % CI | | SD | $d_{average}$ | N |
|---|---|---|---|---|---|---|---|
| | | | Lower | Upper | | | |
| Proficiency | Implicit | 0.56 | 0.35 | 0.77 | 0.53 | | 25 |
| | Explicit | 0.72 | 0.47 | 0.98 | 0.69 | | 29 |
| | Difference | -0.16 | -0.50 | 0.18 | 0.62 | -0.26 | 54 |
| Pronunciation | Implicit | 0.14 | -0.11 | 0.39 | 0.62 | | 25 |
| | Explicit | -0.07 | -0.32 | 0.18 | 0.66 | | 29 |
| | Difference | 0.21 | -0.14 | 0.56 | 0.65 | 0.32 | 54 |
| Discourse | Implicit | 0.52 | 0.31 | 0.73 | 0.53 | | 25 |
| | Explicit | 0.22 | 0.02 | 0.43 | 0.56 | | 29 |
| | Difference | 0.30 | 0.00 | 0.60 | 0.55 | 0.53 | 54 |
| Vocabulary | Implicit | 0.42 | 0.19 | 0.65 | 0.57 | | 25 |
| | Explicit | 0.67 | 0.43 | 0.91 | 0.65 | | 29 |
| | Difference | -0.25 | -0.59 | 0.08 | 0.61 | -0.41 | 54 |
| Grammar | Implicit | 0.40 | 0.20 | 0.60 | 0.50 | | 25 |
| | Explicit | 0.00 | -0.26 | 0.26 | 0.71 | | 29 |
| | Difference | 0.40 | 0.06 | 0.74 | 0.62 | 0.64 | 54 |
| Complexity | Implicit | 0.48 | 0.26 | 0.70 | 0.55 | | 25 |
| | Explicit | 0.57 | 0.29 | 0.85 | 0.75 | | 29 |
| | Difference | -0.09 | -0.45 | 0.28 | 0.67 | -0.13 | 54 |

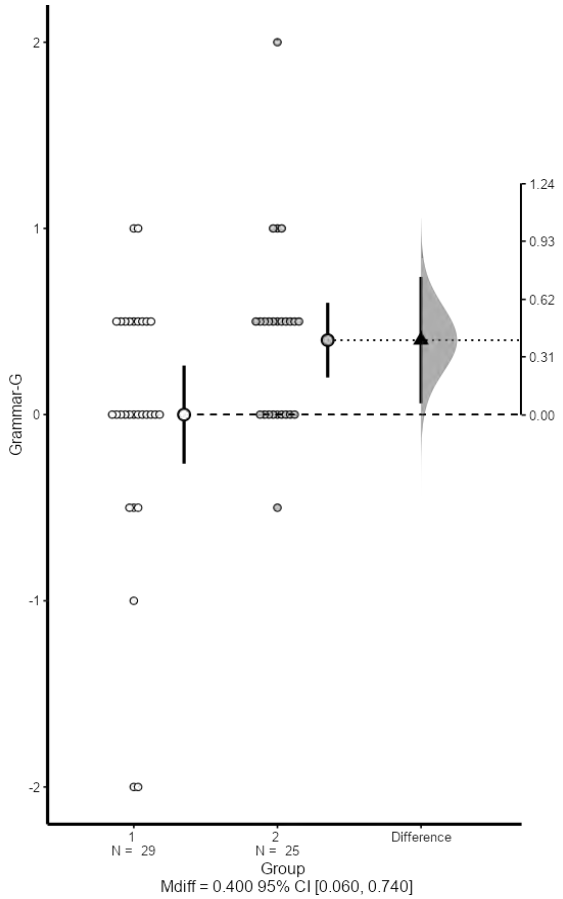
*Note: 1. The standardized effect size was $d_{unbiased}$ because the denominator used was SDpooled. The standardized effect size was corrected for bias. The bias-corrected version of Cohen's d was also called Hedges' g. 2. Negative mean scores in a group indicated that the scores in the post-test were lower than the pre-test.*

.

Table 6 shows that the IG had higher means than the EG for three scales: pronunciation ($M_{diff}$ = -.21, [-.14, .56]), discourse ($M_{diff}$ = .30, [.00, .60]), and grammar ($M_{diff}$ = .40 [.06, .74]). The effect sizes of these three scales ranged from small to medium ($d_{unbiased}$ = .32 for pronunciation, $d_{unbiased}$ = .53 for discourse, and $d_{unbiased}$ = .64 for grammar). Meanwhile, the EG had a higher means than the IG for the other three scales: proficiency, vocabulary, and complexity, although the effect sizes for these were small ($d_{unbiased}$ = -.26 for proficiency, $d_{unbiased}$ = -.41 for vocabulary, and $d_{unbiased}$ = -.13 for complexity). Further investigation was performed to determine the meaning of the difference using a scatter plot for grammar (Figure 4), which had the largest effect size in the analysis.

**Figure 4**

*Scatter Plot of Means and 95% Confidence Intervals (CIs) for the Gain in Grammar Scale for the Two Groups*




Note: 1. The difference between the group means, with its 95% CI, is shown on a floating difference axis at the right. 2. The explicit group is 1, and the implicit group is 2.

Figure 4 shows that for Grammar, the length of the CI on mean difference between the two groups was roughly one and a half times the average of the lengths of the CIs on the two groups' means. This length of the CI is normal for analysis involving two independent groups (Cumming, 2014). These point and interval estimates implied that the result was low for precision since the CI's range was wide. However, Cumming (2014) believes that it could still contribute to a meta-analysis. As can also be seen in Table 6, the IG's CI [.20, .60] was shorter than the EG's [-.26, .26], illustrating that the result in the former group was more accurate than that of the latter.

As shown in Figure 4, the CIs of the two groups overlapped for more than half of one arm, indicating the difference between the two means was not meaningful. These results showed that when measured based on the raters' judgment, both instructional conditions also had a comparable effect on students' oral proficiency.

The explicit strategy training instruction seemed to have more effect on students' general proficiency, vocabulary, and complexity. It seems that instructing participants to explicitly focus on aspects that could improve their speech fluency (such as avoiding and filling their pauses and avoiding making repetitions and repair moves) could improve their oral performance by producing a better-organized speech with varied vocabulary and more complex sentence structures.

Meanwhile, the effect of implicit instruction on participants' pronunciation, discourse, and grammar seemed to be better than explicit instruction. Therefore, it can also be argued that implicit instruction could improve participants' oral performance slightly differently. When instructed implicitly, participants were able to produce a more plausible and logically informed speech. The speech also had fewer phonemic and grammatical errors and better intonation patterns.

## 5. Discussion

The results showed that participants in both groups showed meaningful improvement in their oral proficiency. The effects of both instructional conditions were also found to be comparable. Each instruction seemed to have stronger effects on different aspects of oral proficiency. The explicit strategy training instruction had a better effect on participants' general proficiency, vocabulary, and complexity, while implicit task-based instruction employed in this study seemed to have a better effect than explicit instruction on participants' pronunciation, discourse, and grammar. Therefore, it can be argued that implicit task-based instruction could improve participants' oral performance, albeit slightly differently.

The findings showed that both instructional conditions improved participants' general oral proficiency. These findings corroborate previous research, such as Garbati and Mady (2015), Nergis (2021), Norris and Ortega (2000), and Spada and Tomita (2010), which have reported that classroom instruction significantly affects L2 oral performance of learners. However, most meta-analyses have measured the effects of instruction based on learners' improvement in using specific language features, such as simple and complex forms. No research has been conducted to measure the effects of instruction by using raters' judgments. The current study's findings suggest that the use of raters' judgments might also be necessary for classroom-based studies to measure the effects of instruction on gains in students' oral proficiency. Both instructional conditions in the current study were found to have a meaningful effect on improving students' oral proficiency. Each instruction had a stronger effect on different aspects of oral proficiency. The explicit instruction seemed to highly affect students' general proficiency, vocabulary, and complexity. The implicit instruction also highly influenced students' oral proficiency. Meaningful effects were specifically found on pronunciation, discourse, and grammar.

Previous research has pointed out that instructing learners to explicitly focus on aspects that can improve their speech fluency, such as avoiding and filling

their pauses and avoiding making repetitions and repair moves, could also improve their oral proficiency (Lee, 2019; Rossiter et al., 2010). In the present study, at the end of the training, the students could produce a better-organized speech with varied vocabulary and some embedded clauses. Mostly, 'their skills under explicit strategy training instruction improved by one level from Level 2 (beginning speaking) to Level 3 (competent speaking). However, pronunciation and grammar did not show any improvement after the treatment. It can be assumed that this result was partly due to the effects of the instruction. the students probably focused heavily on making their speech fluent, as this was the focus of the instruction, and somehow forgot to apply correct grammatical codes and accurate pronunciation for the speech.

Previous research has revealed that learners can produce a more plausible and logically informed speech when instructed implicitly (Khayr et al., 2023). In this study, participants' speech had fewer phonemic and grammatical errors and better intonation patterns after they had received implicit task-based instruction. Notably, the skill improvement of participants in this group was still within the same level (Level 3). However, the range of improvement (mean difference between the pre-test and post-test) was almost similar to that of the participants in the explicit strategy training instruction group. This indicated that the implicit task-based instruction group had a slightly higher starting point than the explicit strategy training instruction group in oral proficiency.

The results of the present study suggest that raters' judgments might also be necessary for EFL classroom-based studies to measure the effects of instruction on gains in learners' oral proficiency. EFL classrooms can be heterogeneous in terms of language proficiency; therefore, generalizing the results of an instructional activity is challenging. Standardized tests or measurements used in EFL classroom research might only apply to learners with specific proficiency levels but not to all learners in the same class. EFL teachers could deal

with this heterogeneity because they know their learners and could be sensitive to changes or improvements in learners' proficiency.

## 6. Conclusion

This study found that implicit and explicit instruction significantly improved EFL students' oral proficiency. The teachers, who acted as raters in the present, were able to detect student improvement by using appropriate rating scales. This detection was possible because EFL teachers usually know and understand their learners' language skills, especially from daily classroom interaction. When teachers know their students' language proficiency levels, are sensitive to students' developmental changes due to learning, and are familiar with end-of-course assessments or other regular classroom assessment methods, they then should be able to assess their students' oral proficiency development.

Teachers' subjectivity in assessing or rating learners' performances could be reduced by providing reliable rating scales and training them on how to apply the scales. Teachers could also be informed and assured about the specific purpose of the rating to assess learners' development objectively. In addition, the teaching experience possessed by these teachers offers another advantage in the rating because raters with teaching experience may be more focused on language pragmatics, content, and rhetorical organization than the surface language features (Cumming, 1990; Kim, 2009). These factors would ensure the accuracy and objectivity of the rating results.

Creating an oral performance rating scale could be costly and time-consuming. However, reliable rating scales are readily available. A combination of holistic and analytic ratings should be used (Bachman & Savignon, 1986; Namaziandost & Ahmadi, 2019) to provide each learner's oral proficiency profile. A holistic rating captures the overall impression of oral performance. In contrast, an analytic rating assesses various performance categories, such as content, delivery, organization, and language features. An analytic rubric is used to identify language

subskills, such as grammar, vocabulary, pronunciation, and fluency (Fulcher, 2003; Metruk, 2018). Hence, a combination of a holistic rubric and an analytic rubric could be implemented on research investigating learners' oral performance to have a more comprehensive result.

Overall, the present study found that teachers' judgments could be essential in classroom-based studies investigating L2 oral proficiency. For this purpose, the mixture of a reliable holistic and analytic rubric to rate learners' improvement would be required to obtain valid and detailed information.

## 7. About the Author

Dony Marzuki is an Associate Professor at the English Department of Politeknik Negeri Padang, Indonesia. His research interests include strategy training and instruction, EFL oral fluency and proficiency, explicit and implicit instruction, and teaching speaking. He holds a doctoral degree in Applied Linguistics from Curtin University.

## 8. Acknowledgment

The author would like to acknowledge some people from the School of Education of Curtin University who helped in completing this study such as Dr. Craig Lambert, Dr. Martin Cooper, and Prof Rod Ellis. The study was also enabled by the funding from LPDP Indonesia.

## 9. References

Ahmadian, M. J., & Tavakoli, M. (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research, 15*(1), 35–59. https://doi.org/10.1177/1362168810383329

Ahmadian, M. J., Tavakoli, M., & Vahid Dastjerdi, H. (2015). The combined effects of online planning and task structure on complexity, accuracy and

fluency of L2 speech. *The Language Learning Journal, 43*(1), 41–56. https://doi.org/10.1080/09571736.2012.681795

Bachman, L. F., & Savignon, S. J. (1986). The evaluation of communicative language proficiency: A critique of the ACTFL oral interview. *The Modern Language Journal, 70*(4), 380–390. https://doi.org/10.2307/326817

Bosker, H. R., Pinget, A.-F., Quené, H., Sanders, T., & de Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing, 30*(2), 159–175. https://doi.org/10.1177/0265532212455394

Brevik, L. M. (2019). Explicit reading strategy instruction or daily use of strategies? Studying the teaching of reading comprehension through naturalistic classroom observation in English L2. *Reading and Writing, 32*(9), 2281–2310. https://doi.org/10.1007/s11145-019-09951-w

Brown, J. D. (2001). Pragmatics tests: Different purposes, different tests. In K. R. Rose & G. Kasper (Eds.), *Pragmatics in language teaching* (pp. 301–325). Cambridge University Press. https://doi.org/10.1017/CBO9781139524797.020

Bui, G., Ahmadian, M. J., & Hunter, A.-M. (2019). Spacing effects on repeated L2 task performance. *System, 81*, 1–13. https://doi.org/10.1016/j.system.2018.12.006

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23–48). Pearson Education.

Bygate, M. (2018). *Learning language through task repetition* (Vol. 11). John Benjamins Publishing Company.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201–219. https://doi.org/10.1177/0265532210393704

Choi, I. (2005). Measurability of oral fluency through ASR-based COPI. *Multimedia Language Education, 8*(2), 240–261.

Chou, M. H. (2018). Speaking anxiety and strategy use for learning English as a foreign language in full and partial English - medium instruction contexts. *TESOL Quarterly, 52*(3), 611–633. https://doi.org/10.1002/tesq.455

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62*(1), 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Claridge-Chang, A., & Assam, P. N. (2016). Estimation statistics should replace significance testing. *Nature methods, 13*(2), 108–109. https://doi.org/10.1038/nmeth.3729

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

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31–51. https://doi.org/10.1177/026553229000700104

Cumming, G. (2014). The new statistics: Why and how. *Psychological science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dalman, M., & Plonsky, L. (2022). The effectiveness of second-language listening strategy instruction: A meta-analysis. *Language Teaching Research, 0*(0). https://doi.org/10.1177/13621688211072981

Dao, P. (2020). Effect of interaction strategy instruction on learner engagement in peer interaction. *System, 91*, Article 102244. https://doi.org/10.1016/j.system.2020.102244

de Jong, N., & Perfetti, C. A. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning, 61*(2), 533–568. https://doi.org/10.1111/j.1467-9922.2010.00620.x

DeKeyser, R. M. (1995). Learning second language grammar rules: An experiment with a miniature linguistic system. *Studies in Second Language Acquisition, 17*(3), 379–410. https://doi.org/10.1017/S027226310001425X

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1–16. https://doi.org/10.1017/S0272263197001010

Ellis, R. (2005). *Instructed second language acquisition: A literature review*. Research Division, Ministry of Education.

Ellis, R., Loewen, S., Elder, C., Reinders, H., Erlam, R., & Philp, J. (2009). *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching*. Multilingual Matters. https://doi.org/10.21832/9781847691767

Fathi, J., & Afzali, M. (2020). The effect of second language reading strategy instruction on young Iranian EFL learners' reading comprehension. *International Journal of Instruction, 13*(1), 475–488. https://doi.org/10.29333/iji.2020.13131a

Fathi, J., & Hamidizadeh, R. (2019). The contribution of listening strategy instruction to improving second language listening comprehension: A case of Iranian EFL learners. *International Journal of Instruction, 12*(2), 17–32. https://doi.org/10.29333/iji.2019.1222a

Forbes, K., & Fisher, L. (2018). The impact of expanding advanced level secondary school students' awareness and use of metacognitive learning strategies on confidence and proficiency in foreign language speaking skills. *The Language Learning Journal, 46*(2), 173–185. https://doi.org/10.1080/09571736.2015.1010448

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299–323. https://doi.org/10.1017/S0272263100015047

Fulcher, G. (2003). *Testing second language speaking* (1st ed.). Routledge. https://doi.org/10.4324/9781315837376

Garbati, J. F., & Mady, C. J. (2015). Oral skill development in second languages: A review in search of best practices. *Theory and Practice in Language Studies, 5*(9), 1763–1770. http://dx.doi.org/10.17507/tpls.0509.01

De Graaff, R. & Housen, A. (2009). Investigating the effects and effectiveness of L2 instruction. In *The handbook of language teaching* (pp. 736–755). Blackwell-Wiley. https://doi.org/10.1002/9781444315783.ch38

Hsu, H.-C. (2019). The combined effect of task repetition and post-task transcribing on L2 speaking complexity, accuracy, and fluency. *The Language Learning Journal, 47*(2), 172–187. https://doi.org/10.1080/09571736.2016.1255773

Huang, B., Alegre, A., & Eisenberg, A. (2016). A cross-linguistic investigation of the effect of raters' accent familiarity on speaking assessment. *Language Assessment Quarterly, 13*(1), 25–41. https://doi.org/10.1080/15434303.2015.1134540

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly, 10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Kang, E. Y., Sok, S., & Han, Z. (2019). Thirty-five years of ISLA on form-focused instruction: A meta-analysis. *Language Teaching Research, 23*(4), 428–453. https://doi.org/10.1177/1362168818776671

Khayr, R., Karawani, H., & Banai, K. (2023). Implicit learning and individual differences in speech recognition: an exploratory study. *Frontiers in psychology, 14*, 1–13. https://doi.org/10.3389/fpsyg.2023.1238823

Kim, Y.-H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187–217. https://doi.org/10.1177/0265532208101010

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Longman.

Lambert, C., Aubrey, S., & Leeming, P. (2020). Task preparation and second language speech production. *TESOL Quarterly, 5*(2), 331–365 https://doi.org/10.1002/tesq.598

Lambert, C., Kormos, J., & Minn, D. (2017). Task repetition and second language speech processing. *Studies in Second Language Acquisition, 39*(1), 167–196. https://doi.org/10.1017/S0272263116000085

Lee, S.-C. N. (2019). *The effects of explicit form-focused instruction on L2 oral proficiency development* [Doctoral dissertation]. Temple University Electronic Theses and Dissertations. http://dx.doi.org/10.34944/dspace/1694

Long, M. H., & Robinson, P. (1998). Focus on form: Theory, research, and practice. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 15–41). Cambridge University Press.

Metruk, R. (2018). Comparing holistic and analytic ways of scoring in the assessment of speaking skills. *Journal of Teaching English for Specific and Academic Purposes, 6*(1), 179–189. https://doi.org/10.22190/JTESAP1801179M

Milliner, B., & Dimoski, B. (2021). The effects of a metacognitive intervention on lower-proficiency EFL learners' listening comprehension and listening self-efficacy. *Language Teaching Research, 0*(0). https://doi.org/10.1177/13621688211004646

Mohd Noh, M. F., & Mohd Matore, M. E. E. (2022). Rater severity differences in English language as a second language speaking assessment based on rating experience, training experience, and teaching experience through many-faceted Rasch measurement analysis. *Frontiers in Psychology, 13*, Article 941084. https://doi.org/10.3389/fpsyg.2022.941084

Nakatani, Y. (2005). The effects of awareness - raising training on oral communication strategy use. *The Modern Language Journal, 89*(1), 76–91. https://doi.org/10.1111/j.0026-7902.2005.00266.x

Namaziandost, E., & Ahmadi, S. (2019). The assessment of oral proficiency through holistic and analytic techniques of scoring: A comparative study. *Applied Linguistics Research Journal, 3*(2), 70–82.

Nergis, A. (2021). Can explicit instruction of formulaic sequences enhance L2 oral fluency? *Lingua, 255*, Article 103072. https://doi.org/10.1016/j.lingua.2021.103072

Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta - analysis. *Language Learning, 50*(3), 417–528. https://doi.org/10.1111/0023-8333.00136

Plonsky, L. (2011). The effectiveness of second language strategy instruction: a meta - analysis. *Language Learning, 61*(4), 993–1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x

Reynolds, B. L., Xie, X. S., & Pham, Q. H. P. (2022). Incidental vocabulary acquisition from listening to English teacher education lectures: A case study from Macau higher education. *Frontiers in Psychology, 13*, Article 993445. https://doi.org/10.3389/fpsyg.2022.993445

Rietdijk, S., van Weijen, D., Janssen, T., van den Bergh, H., & Rijlaarsdam, G. (2018). Teaching writing in primary education: Classroom practice, time, teachers' beliefs and skills. *Journal of Educational Psychology, 110*(5), 640–663. https://doi.org/10.1037/edu0000237

Rossiter, M. J., Derwing, T. M., Manimtim, L. G., & Thomson, R. I. (2010). Oral fluency: The neglected component in the communicative language classroom. *Canadian Modern Language Review, 66*(4), 583–606. https://doi.org/10.3138/cmlr.66.4.583

Sato, M. (2020). Metacognitive instruction for collaborative interaction: The process and product of self-regulated learning in the Chilean EFL context. In C. Lambert & R. Oliver (Eds.), *Using tasks in second language teaching: Practice in diverse contexts* (pp. 215–236). Multilingual Matters. https://doi.org/10.21832/9781788929455

Seifoori, Z., & Vahidi, Z. (2012). The impact of fluency strategy training on Iranian EFL learners' speech under online planning conditions. *Language Awareness, 21*(1–2), 101–112. https://doi.org/10.1080/09658416.2011.639894

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta - analysis. *Language Learning, 60*(2), 263–308. https://doi.org/10.1111/j.1467-9922.2010.00562.x

Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning, 71*(2), 285–325. https://doi.org/10.1111/lang.12433

Tavakoli, P., Campbell, C., & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic intervention. *TESOL Quarterly, 50*(2), 447–471. https://doi.org/10.1002/tesq.244

Teng, F., & Huang, J. (2019). Predictive effects of writing strategies for self - regulated learning on secondary school learners' EFL writing proficiency. *TESOL Quarterly, 53*(1), 232–247. https://doi.org/10.1002/tesq.462

Teng, L. S., & Zhang, L. J. (2020). Empowering learners in the second/foreign language classroom: Can self-regulated learning strategies-based writing instruction make a difference? *Journal of Second Language Writing, 48*, Article 100701. https://doi.org/10.1016/j.jslw.2019.100701

Thomas, N. (2020). Incidental L2 vocabulary learning: Recent developments and implications for future research. *Reading in a Foreign Language, 32*(1), 46–60. https://doi.org/10125/66576

Wang, Y. H. (2016). Reading strategy use and comprehension performance of more successful and less successful readers: A think-aloud study. *Educational Sciences: Theory & Practice, 16*(5), 1789–1813.

Wang, Z. (2014). On-line time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 27–62). John Benjamins. https://doi.org/10.1075/tblt.5.02wan

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Xu, Y., Huang, M., Chen, J., & Zhang, Y. (2023). Investigating a shared-dialect effect between raters and candidates in English speaking tests. *Frontiers in Psychology, 14*, Article 1143031. https://doi.org/10.3389/fpsyg.2023.1143031

Ziegler, N., & González-Lloret, M. (2022). *The Routledge handbook of second language acquisition and technology*. Routledge. https://doi.org/10.4324/9781351117586