4/3/2 or 2/3/4? The Impact of Task Design on Ratings of Oral Fluency with Japanese Undergraduate EFL Learners

Rian DAVIS
Michael DELVE
Lydia EBERLY
Craig MERTENS
Thomas STRINGER
Michael WILKINS

Oral fluency is an important language skill and goal for many learners in English as a Foreign Language (EFL) contexts. A commonly used fluency development task is 4/3/2. Learners create a four-minute speech on a topic, then pare it down to a three-, then two-minute speech. Manipulating the language to fit within the shorter timeframe is thought to improve fluency. However, questions have been raised about whether reversing that order might be more effective for EFL learners. This study explores the impact of different models of task design on ratings of EFL learners' oral fluency. The participants were 70 mainly first-year students at a Japanese university. Participants made pretest and posttest treatment recordings of themselves. The treatment was for participants to do either the 4/3/2 or the 2/3/4 speaking fluency activity once a week for four weeks. Each recording was rated at least twice by the team of instructor-researchers to create a pre- and posttest speaking fluency score for each participant. The results of the study were that the treatment did show a significant increase in fluency but that there was not a significant difference between the 4/3/2 treatment and the 2/3/4 treatment. As such, the study represents an important opportunity to re-evaluate the effectiveness of a commonly used teaching tool. Future directions and implications are addressed.

For language learners, speaking fluency, or the ability to speak smoothly, quickly, and with minimal hesitation, is a highly desirable goal. However, for learners studying in an English as a Foreign Language (EFL) context such as Japan, this goal is rarely attained. One explanation for this is the lack of opportunities to use the second language (L2) outside of the classroom (Doe, 2021). Therefore, it is important for EFL learners to gain extensive practice using the target language within the classroom environment. One widely employed activity designed to improve learner fluency is the 4/3/2 technique.

This task was originally described by Maurice (1983) and gained popularity after being advocated to improve oral fluency by Nation (1989), who found that using this activity led to significant gains in participants' rate of speaking. Consequently, a number of studies have been conducted using this task with findings supporting Nation's claims regarding improvements in fluency (e.g. Boers, 2014; De Jong & Perfetti, 2011; Doe, 2021; Thai & Boers 2016).

In the 4/3/2 technique learners are given a few minutes to prepare a monologue on a familiar topic. Next, they give the same monologue to three different partners under shrinking time conditions (four, three, and two minutes). According to Nation (1989), this task has three important characteristics:

1. The speakers have a different listener for each monologue. This is to remove any pressure to add new information and to keep the listener interested during the repetitions.
2. The same monologue is repeated. This gives the speakers easier access to the language needed in the second and third monologues to build fluency. It also develops speaking confidence.
3. The time for each monologue is reduced on each repetition. This also supports fluency and means that new language is not needed to fill the allotted time.

Perhaps the most important characteristic of the 4/3/2 task is repetition. One explanation for this is Levelt's model of speech production (Levelt, 1989 as cited in Boers, 2014). According to this model, fluent speech depends on three elements: selecting the speech content, finding the necessary language to encode this content, and converting the encoded content into speech. When the monologue is repeated, the processing demands involved in choosing speech content and finding necessary linguistic resources are reduced. During the second and third repetitions of the monologue the content has already been decided and the necessary words and grammatical structures should already be available due to prior activation in the first monologue. During the third repetition of the monologue, speakers can focus on and monitor the encoding of their content into speech, which should result in improved fluency (Boers, 2014). De Jong and Perfetti (2011) further argue that repeating monologues enables learners to reuse certain words and structures, and this leads to more efficient processing and higher degrees of automaticity.

In addition to repetition, it has also been found that time pressure contributes to greater fluency in the final speech of the 4/3/2 task. De Jong et al. (2012) and Boers (2014) used two conditions to test the effect of time pressure on fluency. One group repeated three monologues with shrinking time conditions and another group repeated three monologues with constant time conditions. Both studies found that the time pressure group made more fluency gains than the constant time group. However, these studies were

conducted using relatively advanced-level participants in an English as a Second Language (ESL) context. Instructor-researchers (hereafter instructors) involved in the current study have found that for many EFL students in Japan, giving a four-minute speech is a very difficult task and most students struggle to speak continuously in English for this length of time. One possible solution to this problem is to reduce the length of the longest monologue to three minutes, which has been done in studies carried out in similar East Asian contexts (Doe, 2021; Ogawa, 2021; Thai & Boers, 2016). An alternative solution is to use a 2/3/4 format. This would enable students to construct a reasonable two-minute speech and expand it to three and finally four minutes. Although the element of time pressure would be reduced, it was felt that this method might fit the cultural context more effectively whilst maintaining a focus on fluency practice. As far as the instructors are aware, no studies to date have investigated the effects on fluency of reversing the 4/3/2 technique.

For the purposes of the study, fluency was defined in terms of the top rank of the fluency rating scale shown in Appendix A. That is, *fluency means speaking smoothly with hesitations, false starts, or corrections of speech only happening occasionally, and speech is only slightly slower than that of native speakers*.

With a view to adapting the 4/3/2 task design to make it more effective in a Japanese EFL context, the instructors posited the following research questions.

Research Questions

1. Do fluency training tasks have an impact on ratings of fluency?
2. Which task design is more effective at promoting fluency: 4/3/2 or 2/3/4?

METHODS

Participants

Seventy undergraduate EFL students at a Japanese university consented to participate in the research, completed all measures correctly, and were included in the study. Participants came from eight classes. Two participants were enrolled in a two-credit *Speaking & Listening* course, and all others were in the university's three-credit *Intensive English* course. These courses share common attainment objectives related to developing speaking skills including fluency. Language proficiency information was only available for IE participants. Their class average TOEIC scores ranged from 428.9 to 630.3. The whole sample had a mean age of 18.97 years (SD = .87) with ages ranging from 18-22. The whole sample was 68.6% female. Two participants from the 2/3/4 treatment group were Chinese, and one from the 4/3/2 group was from South Korea. The remaining 95.7% of participants were Japanese. See Table 1 for detailed sample demographic information of the 2/3/4 and 4/3/2/ treatment

groups. Participants read and signed consent forms written in Japanese, and were debriefed after the end of the data analysis.

TABLE 1
Sample demographics by treatment group

Treatment Group	Frequency (*n*)	Percent (*%*)	Mean Age (*SD*)	Gender (*Male/Female/ Other*)
2/3/4	40	57.14	19.1 (.90)	12/28/0
4/3/2	30	42.86	18.8 (.81)	9/20/1

Measures

Fluency Rating Scale

Rating scales are a commonly used tool in language testing studies to help expert raters assess the performance of test-takers (O'Grady, 2019). Appendix A shows a five-stage fluency rating scale that was adopted for use in the study from Doe (2021). The scale ranges from 1, '*Speech is so halting and fragmentary that conversation is impossible*' to 5, '*Speaks fairly fluently with only occasional hesitation, false starts, and modification of attempted utterance. Speech is only slightly slower than that of a native speaker*.' The scale has previously been shown to reliably assess English fluency in Japanese university contexts (Doe, 2021; Nitta & Nakatsuhara, 2014). Indeed, Iwashita et al. (2001), the original developers of the scale, demonstrated its reliability at assessing English language fluency among a broad range of L1 speakers of Asian languages including Japanese.

Treatment and Recordings: Speaking Prompts

In the 2/3/4 treatment group, participants made monologue speeches about a selection of speaking prompts for two minutes, then three minutes, then four minutes. Conversely, in the 4/3/2 treatment group, participants spoke for four minutes, three minutes, then two minutes. This was done to allow assessment of the impact of different task designs on speaking fluency. Additionally, in the pre- and posttests, participants made similar speeches in response to prompts. Participants did not do 4/3/2 or 2/3/4 activities in these test weeks. For these elements of the study, a series of 11 speaking prompts was selected. These prompts were drawn from past papers of the IELTS English proficiency exam, Speaking section, Part 2. Part 2, known as the *individual long turn*, uses

prompts of general interest on familiar topics (Seedhouse & Harris, 2011) and has a similar procedure to this study; presentation of a speaking prompt, allotted preparation time, and a monologue speech for up to two minutes made to another person (Iwashita & Vasquez, 2015). IELTS speaking prompts are open-ended, and allow respondents with a wide range of proficiencies to showcase their abilities. As such, they were used in the study. All prompts selected for the study began in a consistent manner, by asking participants to describe something. This was followed by a few content suggestions for the speech. For instance, the practice speaking topic was '*Describe a book that you enjoyed reading because you had to think a lot*' and the content suggestions were to talk about '*the name of the book', 'why you decided to read it', 'what the book made you think about'* and '*why you enjoyed reading this book'*. For a full list of speaking prompts in the context of the research schedule, see Table 2.

Study Procedures

The study adopted a between-groups design. Intact classes were assigned to either the 2/3/4 (n = 40) or 4/3/2 (n = 30) treatment group; however, there was a group imbalance, as shown in Table 1 and discussed in the Limitations section of this paper. Inclusion criteria were: giving informed consent, being absent no more than once during the treatment or at all during data collection, and producing ratable speaking recordings. In the study, all participants completed all measures. Although the addition of a control group might increase the strength of potential findings, it was decided that the ethical risks of withholding a potentially beneficial pedagogical tool from some participants outweighed the marginal research benefits. Data was collected over a six-week period between November and December 2022. Pre- and post-test monologue speaking recordings were made and collected using an online platform, *ZenGengo* (www.zengengo.com), and participants' personal smart devices. Participants took different courses with different scheduling. *Intensive English* meets three times a week and *Speaking & Listening* meets twice a week, and all class meetings are for 100 minutes. However, as shown in the research schedule in Table 2, all recordings were collected consistently. Recording took place as part of in-class activities in weeks one and six, before and after the treatment. In week one, before the pre-test recording was collected, all participants completed a practice session in class. Participants were shown a speaking prompt on the projector and instructed to prepare written notes, thinking about what they wanted to say without the assistance of their smart device or a dictionary. Previous research in this area has demonstrated fluency gains with allotted preparation time of as little as one (Doe, 2021) or as much as five minutes (De Jong & Perfetti, 2011).

TABLE 2
Research Schedule

		Strand 1 (234; AB)	Strand 2 (234; BA)	Strand 3 (432; AB)	Strand 4 (432; BA)
Week 1	Practice	*Describe a book that you enjoyed reading because you had to think a lot*			
	Pretest	A	B	A	B
Week 2	Treatment Session 1	*Describe a hotel you know*		*Describe a tourist attraction you enjoy visiting*	
Week 3	Treatment Session 2	*Describe a famous businessperson that you know*		*Describe a review you read about a product or service*	
Week 4	Treatment Session 3	*Describe something you liked very much which you bought for your home*		*Describe a luxury item you would like to own in the future*	
Week 5	Treatment Session 4	*Describe a very difficult task that you succeeded in doing as part of your work or studies*		*Describe some technology that you decided to stop using*	
Week 6	Posttest	B	A	B	A

Note. Prompt A was '*Describe a website you bought something from*', and Prompt B was '*Describe an interesting TV programme you watched about a science topic*'

In the study, five minutes of preparation time were allotted to balance available class time and allow sufficient time for preparation. Instruction was then given on how to make a good-quality recording. For instance, participants were encouraged to hold the microphone of their smart devices at a consistent distance from themselves throughout the recording. At this point, participants were also reminded that the research activities were not graded, to reduce any potential anxiety. Finally, each participant scanned a QR code displayed from the projector using their device to access *ZenGengo* (www.zengengo.com) and make their recording. At this stage, the instructors provided support for participants who had technical issues. Having been familiarized with the task and recording procedures, participants repeated the process to generate the pretest recordings with new prompts in the same class period.

Classes had been grouped into four strands by treatment group, 2/3/4 or 4/3/2, and also by pre- or posttest speaking prompt order, A or B. As shown in Table 2, Strands 1 (2/3/4) and 3 (4/3/2) were given prompt A as a pretest, whereas Strands 2 (2/3/4) and 4 (4/3/2) were given prompt B. Prompts were varied to avoid any potential practice effect on the results.

All classes then participated in the treatment program, which occurred in once-weekly in-class sessions. Previous studies in this area have noted fluency increases with one (Boers, 2014), three (De Jong & Perfetti, 2011), and 11 treatment sessions (Ogawa, 2021). As available class time was limited, four successive weekly treatment sessions were held between weeks two and five. The treatment was administered by the instructors and as shown in Table 2, consisted of speaking prompts that differed between the study strands. Participants followed the same procedures as in previous stages, except rather than making a recording they discussed the same prompt with a partner in three successive rounds, changing partners between rounds. In the 2/3/4 treatment group, these rounds increased in length by one-minute increments per round. In the 4/3/2 treatment group, the rounds decreased in length by one-minute increments per round. Finally, an in-class post-test recording was collected in week 6. Although a longer treatment period would have been preferable, this was all that was possible within the available time. As seen in Table 2, the order of speaking prompts A and B were reversed for each strand at posttest. This was done to mitigate the impact of inconsistencies in speaking prompt difficulty on the analysis, as in Doe (2021).

Data Rating Procedures

The data for this study were 140 paired monologue speaking recordings made by the 70 participants. Raters reviewed and rated each recording according to the fluency rating scale. The raters were the six instructors involved with the study, who all hold advanced degrees in language education or related fields and have many years of experience in grading speech samples from EFL students using rubrics. Various study design procedures were undertaken to enhance the reliability of their ratings. One such measure was

sample anonymization. First, participants were all assigned a random identifier. Second, recordings were then assigned a suffix denoting an A or B speaking prompt sample. Recording files were then renamed accordingly, for example, P01_A or P02_B. As the order of A and B speaking prompts alternated between strands, this meant raters could not identify the participant or whether a recording was from a pre- or posttest sample. This indicated their rating might be less susceptible to bias.

Other measures were undertaken to enhance inter-rater reliability and the consistent application of the fluency rating scale by all raters. An approximately 20-minute rater training meeting was held, at which four randomly selected recordings from the pretest samples were played and analyzed. Previous scholarship in this area identified that rater training reduced variability and promoted reliability when expert raters graded EFL learners' performance on the IELTS Speaking section (Doosti & Safa, 2021). In the study, the rating scale was first introduced at the meeting. All six raters privately rated the four recordings then discussed ratings and reasons. No rating discrepancies of more than one level on the fluency scale were noted. Furthermore, at the meeting, it was decided that all six instructors would rate, and that two raters would review and rate each of the 140 recordings as an additional reliability-enhancement measure. Other researchers in this field used this approach when evaluating audio samples of IELTS speaking performance for assessment (Nakatsuhara et al., 2021). In the study, if both raters gave the same score, then that score was assigned. Discrepancies of one level on the fluency scale were resolved by averaging their scores. Furthermore, cases of inter-rater discrepancy greater than one level of the fluency scale were resolved by a third rater being assigned. In this study, this happened in only two percent of cases, implying that the measures adopted had been broadly successful and raters assessed the recordings with acceptable consistency.

Finally, rater assignments were randomized. Efforts were made to avoid the same rater listening to both samples from a single participant; however, this was not always possible. Raters then independently listened to between 40 and 50 recordings each. Only the first two minutes of each recording were listened to, as this was the lower and upper bound of the 2/3/4 and 4/3/2 groups, respectively. Another reason was to avoid prejudicing their ratings by listening to speakers who spoke for significantly longer than two minutes. This approach also mirrors the test procedure in the IELTS Speaking section Part 2 from which the prompts were adopted. Recordings where the speaker failed to speak for two minutes were noted and will be discussed in the Limitatoins section. Finally, raters compiled their ratings on a shared spreadsheet for quantitative statistical analysis.

RESULTS

As the dependent variable was ordinal, non-parametric tests were used. A Wilcoxon signed-rank test showed that in aggregate, the posttest scores (Mdn = 3.5) were significantly higher than the pretest scores (Mdn = 3), W = 355.5, p = .013, as seen in Appendix B.

The Mann-Whitney U Test was used to compare the two independent treatment groups. There was a significant difference between the two groups at both the time of the pretest (U = 820, p = .009) as well as at the time of the posttest (U = 773, p = .042). At both times the participants in the 2/3/4 group had higher median scores than those in the 4/3/2 group. Appendices C and D show the box plot distributions by treatment group.

When comparing the difference between each participant's pretest and posttest scores (Appendix E), there were no significant differences between the treatment groups (U = 552.5, p = .530). The 2/3/4 group's mean increase was 0.15 points, and the 4/3/2 group's mean increase was 0.29 points.

Comparing the two different prompts used for the tests, A and B, a Wilcoxon signed rank test showed no significant differences between them (W = 644.5, p = .548).

DISCUSSION

In this study an oral fluency development activity was employed in EFL classes at a Japanese university with two task designs: 4/3/2 and 2/3/4. The study focused on two questions: Did fluency training have an impact on ratings of fluency and which task design was more effective at promoting fluency? The activity was performed once a week over a four-week period, and speaking recordings were collected in a pretest and posttest arrangement. The instructors used a fluency rating scale and recorded their ratings for analysis. The analysis showed that the treatment led to statistically significant fluency gains. However, there was no significant difference between the task orders (4/3/2 or 2/3/4) on fluency. This is a novel finding. As such, the study both confirms previous research findings on the 4/3/2 speaking activity (e.g. Boers, 2014; De Jong & Perfetti, 2011; Doe, 2021; Thai & Boers 2016), while also casting new light on the exact mechanism of change. Pedagogical and research implications, in addition to limitations, are discussed below.

Design Effectiveness

In this study, both the 4/3/2 and 2/3/4 activities were equally effective at promoting fluency. In addition to the two different task treatments, the results show the type of topic used for data collection did not have an impact either way on ratings of participants' fluency. However, the 2/3/4 group did have a higher rating in both the pretest, seen in Appendix C, and the posttest, seen in Appendix D, than the 4/3/2 group. This could be attributed to random variation. Another possibility is that perhaps participants were intimidated by the initial length of the first stage of the activity. Participants in the 2/3/4 group had to

start by speaking for only two minutes, while the 4/3/2 group had to start out by speaking for a full four minutes. For lower intermediate participants in the 4/3/2 treatment group, the four-minute starting time might have increased task anxiety more than those in the 2/3/4 group. However, the study did not measure the interaction of anxiety or proficiency on fluency.

Fluency Training

The aggregate fluency ratings of both 4/3/2 and 2/3/4 at the end of the study period showed an increase (see Appendix C). However, the order of the activity, being 4/3/2 or 2/3/4, showed no significance. Therefore, these results suggest that repetition (Levelt, 1989 as cited in Boers, 2014), or a combination of repetition and some change in time allowance (be it increasing or decreasing), rather than the increasing pressure alone, might be the key mechanism of change in fluency rating over a period of time. The activity was performed four times; however, the repetition of the speaking content was performed three times each. For future consideration, a speaking activity using the same amount of time, three minutes each for three times in a row, compared to a 4/3/2 or 2/3/4 speaking activity may be of interest to see if time change is necessary at all. The researchers in the Thai and Boers (2016) study compared 4/3/2 to constant time task repetition and found little benefit in 4/3/2 over other patterns.

Motivation

Motivation may also have been a factor in fluency increases reported here. After completion of the first stage of the treatment, there may have been differences in motivation to complete the remaining task stages between treatment groups. If a participant spoke for two full minutes but no more, then the participant in the 2/3/4 group may have had higher motivation to complete the activity having done the first stage correctly and try hard to finish the activity well, whereas a participant in the 4/3/2 group may feel demotivated by not fully completing the first stage of the activity andput forth less effort in the final stages. However, the study did not measure the impact of motivation on fluency.

LIMITATIONS

Rating Scale

A factor that might have affected the results is a possible omission in the rating scale itself. Although several parameters to rate fluency are given by the scale, as shown in Table 2, it does not clearly state the total length of speaking time in any stage of the scale. If a minimum duration had been included in the rating scale the instructors could have taken certain instructional measures or made clearer and more concrete decisions while rating the recordings. Also, this rating system only focuses on limited aspects of fluency and not on other parts of speaking, such as the use of difficult grammar or non-frequent vocabulary. Would a student's recording of more complex language use with pauses and

errors be considered less fluent than another participant using simpler vocabulary and language structures in a clearer and quicker manner?

Pretest and Posttest Speaking Recordings

Some technical issues also affected the speaking recordings. For example, some participants' audio was distorted or only audible in short spurts, voices were not clear, or the entire audio recording was completely silent for the entire duration. These issues were limited in numbers (12), but the most common issue was length of the recording. The instructions for the pretest and posttest stated that the length of the recordings should be up to four minutes; however, a large number (44) of participants finished their audio recordings significantly under the four-minute mark, ranging from 30 seconds to somewhere under two minutes. This created a question of whether to include these recordings in data collection or omit them. Ultimately, it was decided to include all audible recordings and their ratings in the data analysis regardless of length. Even with short recordings, raters were able to assess fluency.

Study Attrition

Several factors that were beyond the control of the instructors could have influenced the results. These include the varying levels of ability and motivation between participants in different classes. In addition, the participants could be taking different types of English classes along with the ones in the study. Another factor is that many students completed the activities, including the pre- and posttest, but did not give consent for their data to be included in the analysis. A final factor relates to participants who were absent from multiple sessions and were excluded from the study. These limitations along with observations in fluency training and design effectiveness should be considered and are described further in the next section on future directions of study.

FUTURE RESEARCH

Longer Research Periods

The instructors who carried out the research felt that there was not enough time for the participants to benefit from the 4/3/2 technique. Perhaps if the training had been spread out over a semester's length or even two semesters rather than roughly a month in which the study was carried out, a larger effect would have been detected.

Use of Qualitative Data

Qualitative data relating to participants' perceptions of task difficulty, topic selection, and fluency gains was collected but its analysis was beyond the scope of the study. Future research will address the results of these surveys. For instance, did the participants perceive a difference after carrying out the tasks? How did they feel about the order of the treatment tasks, whether 4/3/2 or 2/3/4? These are just a few examples of survey questions that would help explain the data and guide instructors in future research directions.

Different Fluency Measures

One of the issues facing the instructors was that the speaking data from the participants did not fit into a 'one-size-fits-all' category. To remedy this, more fluency metrics could be incorporated into the research such as the length of speech, words-per-minute, number of pauses and other data.

Investigating Participant Specific Factors

Another consideration is the data on the participants themselves and how factors such as target language proficiency may or may not affect measurable improvement in participant fluency. Did more proficient participants fare better or worse in terms of measured fluency improvement? Or did language proficiency not matter at all in terms of meaningful improvement? These are important questions that remain unanswered. In the present study, only class average TOEIC scores were available, rather than individual participant's standardized proficiency data. Other factors for consideration are the demographic data of the participants such as age. Whether or not the participant considers him or herself outgoing or not may be another interesting direction to investigate.

Use of Alternative Rating Scales

The rubric used is widely tested, but it may not be appropriate for this task. Perhaps other rubrics or ways to rate the fluency would achieve different results. A fluency scale that separates participant data into finer categories might have been useful and should be considered in future research. For instance, categories that detect more nuances in the fluency rating as compared to native speakers when grading the participants' speech output might have yielded more insightful data. Furthermore, due to the many varieties of Englishes spoken worldwide, each with equal validity, views on what can be regarded as 'native speaker' level may be too subjective to definitely rank the participants' utterances. To solve this issue, more objective criteria could be used without referring to 'native speaker level' in the rubric.

CONCLUSION

In recent years, fluency has become a salient topic. Overall, the results of this study indicated that 1) there was a relationship between the training and an increase in fluency among the participants; however, 2) it did not matter whether the practice speeches were decreasing or increasing in length (4/3/2 or 2/3/4). Despite these mixed results, new light has been thrown on mechanisms of change in fluency, and several possible ways to continue research in regards to the 4/3/2 technique have been noted.

REFERENCES

Boers, F. (2014). A reappraisal of the 4/3/2 activity. *RELC Journal*, *45*(3), 221-235. https://doi.org/10.1177/0033688214546964

De Jong, C.A.M. (2012). Does time pressure help or hinder oral fluency? In N. De Jong, K. Juffermans, M. Keijzer, & L. Rasier (Eds.), Papers of the Anéla 2012 Applied Linguistics Conference (pp. 43-52). Eburon. DeJong_Anela_2012

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

De Jong, N., Florjn, A., Hulstijn, J., Schoonen, R., & Steinel, M. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*(1), 5–34. https://doi.org/10.1017/S0272263111000489

Doe, T. (2021). Fluency development in an EFL setting: A one-semester study. *Language Teaching Research*, *0*(0), 13621688211058520. https://doi.org/10.1177/13621688211058520

Doosti, M., & Safa, M. A. (2021). Fairness in oral language assessment: Training raters and considering examinees' expectations. *International Journal of Language Testing*, *11*(2), 64. https://files.eric.ed.gov/fulltext/EJ1318848.pdf

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency Test? Exploring the potential of an information-processing approach to task design. *Language Learning*, *51*(3), 401-436. https://doi.org/https://doi.org/10.1111/0023-8333.00160

Iwashita, N., & Vasquez, C. (2015). *An examination of discourse competence at different proficiency levels in IELTS Speaking Part 2* (IELTS Research Report Series, Issue ISSN 2201-2982). https://www.ielts.org/-/media/research-reports/ielts_online_rr_2015-5.ashx

Maurice, K. (1983). The Fluency Workshop. TESOL Newsletter 17, 429. https://eric.ed.gov/?id=ED289347

Nakatsuhara, F., Inoue, C., & Taylor, L. (2021). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*, *18*(2), 83-106. https://doi.org/10.1080/15434303.2020.1799222

Nation, I.S.P. (1989). Improving speaking fluency. *System*, *17*, 377-384. https://doi.org/10.1016/0346-251X(89)90010-9

Nitta, R., & Nakatsuhara, F. (2014) A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, 31(2), 147-175. https://doi.org/10.1177/0265532213514401

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English-medium university admission. *Language Testing*, *36*(4), 505-526. https://doi.org/10.1177/0265532219826604

Ogawa, C. (2021). Revised 4/3/2 task: Fluency training with formulaic language in the EFL classroom. *The Journal of Asia TEFL*, *18*(4), 1108-1127. https://doi.org/http://dx.doi.org/10.18823/asiatefl.2021.18.4.3.1108

Seedhouse, P., & Harris, A. (2011). *Topic development in the IELTS Speaking Test.* IELTS RESEARCH REPORTS, 12(1),1-44 . https://www.ielts.org/-/media/research-reports/ielts_rr_volume12_report2.ashx

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects on fluency, complexity and accuracy. *TESOL Quarterly*, 50, 447-471. https://doi.org/10.1002/tesq.232

APPENDIX A: Fluency Scale

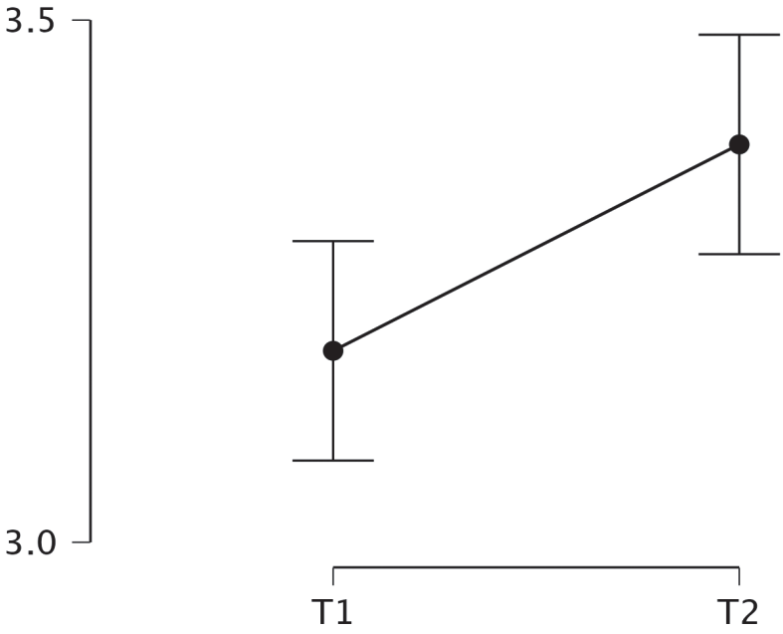
The fluency rating scale used in the study

5	Speaks fairly fluently with only occasional hesitation, false starts and modification of attempted utterance. Speech is only slightly slower than that of a native speaker
4	Speaks more slowly than a native speaker due to hesitations and word-finding delays
3	A marked degree of hesitation due to word-finding delays or inability to phrase utterances easily
2	Speech is quite disfluent due to frequent and lengthy hesitations or false starts
1	Speech is so halting and fragmentary that conversation is impossible

Source. Scale adopted from Doe (2021)

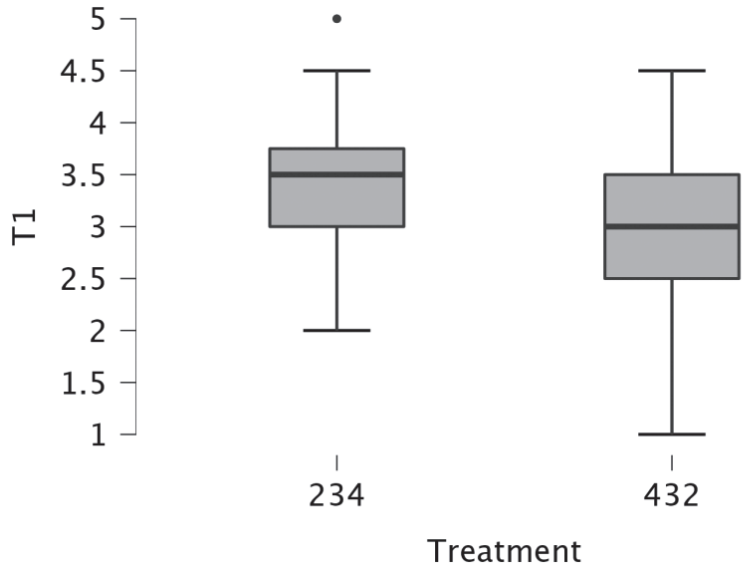
APPENDIX B: Results Comparing Pre- and Post-test Scores

Line graph comparing the pre-test (T1) and post-test (T2) scores.



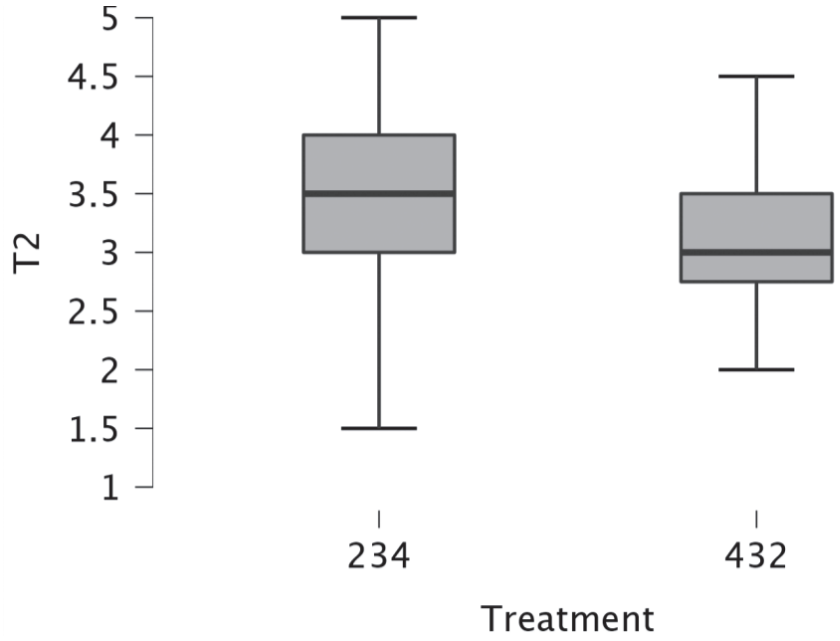
APPENDIX C: Result of Pre-test Scores by Treatment

Box plot of distribution of pretest scores by treatment



APPENDIX D: Results of Post-test Scores by Treatment

Box plot of distribution of post-test scores by treatment



APPENDIX E: Results of Gains by Treatment

Box plot comparing distribution of gains by treatment