TESL-EJ
Celebrating Issue #100
Teaching English as a Second or Foreign Language

The Electronic Journal for English as a Second Language

The Effects of L2 Pronunciation Instruction on EFL Learners' Intelligibility and Fluency in Spontaneous Speech

February 2022 – Volume 25, Number 4

Tuc Chau
University of South Florida
<tuccaochau@usf.edu >

Amanda Huensch
University of Pittsburgh
<amanda.huensch@pitt.edu>

Yen K. Hoang
Gia Viet English Language Center
<hoangkimyen0309@gmail.com>

Hiep T. Chau
Can Tho University
<chiep@ctu.edu.vn>

Abstract

This study investigated the effects of L2 pronunciation instruction on speech intelligibility and fluency, the relationship between intelligibility and fluency, and the extent to which utterance fluency can predict perceived fluency. Participants were 30 beginning adult EFL learners who received either segmental or suprasegmental instruction. Oral data included monologues recorded at the beginning and end of an 8-week course. Speech segments were transcribed for intelligibility and rated on a 1000-point scale for fluency by 11 native speakers. They were also coded and analyzed for transcription errors and utterance fluency measures. Quantitative analyses did not reveal significant changes in intelligibility or perceived fluency as the result of instruction. However, the suprasegmental group seemed to show an upward trend in speech rate, which was found to strongly predict perceived fluency. The findings further our understanding of the effectiveness of pronunciation instruction based on an understudied population and a free response outcome measure.

Keywords: *pronunciation, intelligibility, direct instruction*

For years, the global adoption of English as an indispensable means of communication for speakers of different L1s has created enormous demands on the teaching and learning of the language as an L2. More than that, English has assumed a privileged position of lingua franca, displaying varying norms across its many contexts of usage (see Jenkins et al., 2011). As a result, the last two decades have witnessed great strides toward L2 pronunciation instruction goals that prioritize intelligibility over nativeness (e.g., Jenkins, 2000; Jenkins, 2007; Levis, 2005; Munro & Derwing, 2015; Murphy, 2014; Seidlhofer, 2013). Nevertheless, recent systematic reviews of pronunciation instruction studies have indicated that research in this area relies heavily on the nativeness principle (i.e., that achieving native-like pronunciation is both possible and desirable, Levis, 2005) when it comes to the methodological decisions employed to measure pronunciation improvement. For instance, Thomson and Derwing (2015) reported that 75% of the studies included in their narrative review evaluated pronunciation performance by focusing on discrete pronunciation features such as the accuracy of individual phonemes instead of using global measures of speech (e.g., intelligibility transcriptions and comprehensibility ratings). Relatedly, these reviews have indicated that much pronunciation instruction research relies on speech samples elicited using controlled (e.g., reading aloud, sentence imitation) rather than spontaneous speech (Lee, Jang, & Plonsky, 2015; Thomson & Derwing, 2015). Thus, while these reviews have indicated that pronunciation instruction is generally effective at improving discrete features in controlled tasks, less is known about the effects of instruction on improving global features in spontaneous speech.

In addressing some of these limitations, the present study contributes to the literature in the following ways. First, it reports on intelligibility (extent of understanding; operationalized as number of words transcribed correctly from speech) and fluency (operationalized as both temporal measures and listener judgements of speech). In focusing on the effects pronunciation instruction has on these global speech dimensions rather than individual phonemes, the design reflects current teaching goals moving away from nativeness principles. It is also worth noting that intelligibility was chosen over comprehensibility because it is relatively understudied, and it is closely related to comprehensibility (Derwing & Munro, 1997; Munro & Derwing, 1995). Accentedness was not included as it is arguably no longer a priority compared to intelligibility within current L2 pronunciation frameworks. Fluency was included to probe its relationship with intelligibility, which remains under-researched (Thomson, 2015). Second, the study applies a spontaneous speaking task (i.e., monologues) rather than relying on a read-aloud task to assess learners' speech, to better understand the effects pronunciation instruction has on spontaneous speech. Exploring potential gains in spontaneous, as opposed to controlled, speech is important because speaking spontaneously in daily communication is likely the goal for many learners. In addition to addressing these gaps, the study provides evidence from the context of Vietnam, where there is a dearth of research on how pronunciation instruction can affect learner speech. Pronunciation instruction research with EFL Vietnamese learner participants allows us to see whether previous findings can be reproduced in this context and the extent to which common pronunciation instruction practices in Vietnamese classrooms are effective.

Literature Review

The Effectiveness of Pronunciation Instruction

Research on the effects of pronunciation instruction has demonstrated that pronunciation instruction is generally effective. In a review of pronunciation instruction effects, Saito (2012) examined 15 quasi-experimental studies that employed pretest-posttest designs to investigate the effects of instruction on L2 pronunciation development. The results indicated that pronunciation instruction is effective regardless of the focus of instruction (i.e., segmental- or suprasegmental-based); however, the evidence of positive effects was primarily obtained from controlled outcome measures (e.g., reading-aloud, elicited imitation). Such measures may guarantee the production of target forms but may not accurately portray a learner's ability in communicative contexts given the greater demands of speaking spontaneously. In a similar vein, Lee et al. (2015) and Thomson and Derwing (2015) conducted large-scale syntheses and analyses of empirical studies to help determine the overall effectiveness of pronunciation instruction, and importantly, to explore potential causes and moderators of effect variance. Lee et al. meta-analyzed a total of 86 studies and found that pronunciation instruction has statistically large effects, but that different contextual (e.g., target languages, institutional types, instructional settings, proficiency levels) and methodological (e.g., treatment types, outcome measures) variables impacted the size of the effects. Similar to Saito (2012), they reported larger effects for read-aloud tasks compared to less controlled outcome measures such as picture narratives or monologues. The results also indicated larger effects for second language contexts (i.e., where the target language is spoken in the local community) in comparison to foreign language contexts. Thomson and Derwing took a narrative approach to reviewing 75 studies, many of which were included in Lee et al.'s meta-analysis and similarly indicated that a majority of the studies (82%) displayed a positive impact on the participants' pronunciation.

Nevertheless, in their accounts, both Lee et al. (2015) and Thomson and Derwing (2015) remain cautious of these encouraging results as they report several methodological issues that need to be addressed. First, both studies observed that only a limited subset of pronunciation features had been considered for intervention, and Thomson and Derwing pointed out that whether instruction led to better speech as a whole was unclear. In other words, empirical evidence of improved intelligibility as a result of pronunciation instruction is still limited (Thomson & Derwing, 2015). To address this limitation, the current study explores to what extent instruction leads to more intelligible and fluent speech rather than focusing on improving discrete features.

Second, concerning outcome measures, pronunciation instruction research is quite dependent on controlled tasks (e.g., reading-aloud, elicited imitation), and thus confines its external validity (Lee et al., 2015). A greater variety of outcome measures, especially those that reflect spontaneous communication, are needed so that evidence of improvement can be interpreted as applicable to real-world contexts (Thomson & Derwing, 2015). One common method to elicit more spontaneous speech in L2 pronunciation literature is with picture description tasks, sometimes also called narrative tasks. In their longitudinal study, Derwing and Munro (2013) used one such narrative task to evaluate ESL learners' comprehensibility, fluency, and accentedness. The learners' narratives, which were based on an 8-frame cartoon story, were recorded and used as stimuli which were rated by both native and nonnative speakers. Although the task is extemporaneous, Derwing et al. (2004) acknowledged that picture narratives impose ideas on L2 students, which may also require certain lexical items and grammatical structures. Given this limitation of narratives, but with a goal of incorporating a less controlled speaking task, the current study assessed the improvement by employing monologues of answers to questions about everyday life, which offered the learners more freedom regarding what they would say and how they would communicate it.

Another aspect of study design to consider relates to which features of pronunciation should be prioritized for pronunciation instruction to yield the most fruitful results. Pronunciation instruction is usually categorized as focusing on segmentals (discrete sounds, i.e., consonants and vowels) or suprasegmentals (units extending beyond individual sounds, e.g., rhythm, stress). The commonly accepted view is that teachers should not intentionally aim at one and neglect the other because both segmental and suprasegmental errors can cause problems for communication (Celce-Murcia et al., 2010). It is thus not the goal of the current study to determine which type of instruction is “best” when both instructional approaches are considered, but rather to build upon the limited number of empirical studies that have used a similar design to increase the validity of previous results. Derwing et al.’s (1998) is, for example, one of the few studies which attempted to investigate the effects of pronunciation instruction on global speech dimensions. They examined the effects of three types of instruction (segmental, global speaking habits and prosodic factors, and no specific pronunciation instruction) on the speech of 48 intermediate adult ESL students in Canada. Before and after the treatment, the students read simple statements aloud and completed a picture narrative task. Listener ratings indicated that for sentence reading, both the experimental groups showed improvement in comprehensibility whereas for the narratives, only the global group showed improvement in comprehensibility and fluency. These results raise important questions about the transferability of pronunciation instruction effects to spontaneous speech. Additionally, because it was not included as a measure, it is unclear how pronunciation instruction might have impacted intelligibility. Thus, the current study employed a similar design but, importantly, included an intelligibility measure.

Intelligibility

According to Levis (2005, 2020), there have been two fundamental principles guiding pronunciation instruction: the nativeness principle and the intelligibility principle. While the former encourages the acquisition of native-like speech, the latter emphasizes the goal of being understandable and rejects goals of eliminating foreign accent. Following nativeness principles can be problematized on multiple levels. For instance, the nativeness principle is insensitive to contexts where L2 learners communicate with each other rather than native speakers (Levis, 2005). Moreover, there is empirical evidence suggesting that adult learners do not typically achieve native-like proficiency in the L2 (Abrahamsson & Hyltenstam, 2009), but that even heavily accented speech can be intelligible (Munro & Derwing, 1995). Hence, the focus of pronunciation research and practice has shifted to improving intelligibility as opposed to pursuing native-like speech.

Broadly, intelligibility is described as “the extent to which a speaker’s message is actually understood by a listener” (Munro & Derwing, 1995, p. 76). One common way in which it can be assessed is calculating the percent of words correctly transcribed by listeners. As an example, Munro and Derwing (1995) asked native speakers to listen to and transcribe excerpts of nonnative speakers’ narratives of a picture-based story. Following the same procedure, Parlak (2010) had raters transcribe short speech samples in standard orthography after listening to them once. In each case, the transcriptions were coded for exact word match, and intelligibility scores were calculated based on the discrepancies between the raters’ transcriptions and the researchers’. The current study adopted this approach.

Fluency

Fluency, in addition to complexity and accuracy, is one of the core constructs of L2 proficiency (Housen et al., 2012), and perceived fluency is a component of global pronunciation proficiency (Saito & Plonsky, 2019). In the narrow sense, fluency refers to the fluidity with which language is spoken (Thomson, 2015). Segalowitz (2010) conceptualized fluency in terms of cognitive fluency (efficiency of cognitive processes involved in producing speech), utterance fluency (temporal measures of speech), and perceived fluency (listener interpretations of cognitive fluency based on utterance fluency). Compared to cognitive fluency, utterance fluency and perceived fluency tend to be more widely adopted in L2 research due to their operationalizability. Utterance fluency can be divided into three subdimensions of speed fluency, breakdown fluency, and repair fluency (Tavakoli & Skehan, 2005). Such categorization is not without problems. Many researchers have argued that speech rate and mean length of run, as operationalizations of speed fluency, might be better categorized as composite measures because they can capture both the speed and breakdown fluency dimensions (Bosker et al., 2013; De Jong et al., 2013; Huensch & Tracy-Ventura, 2017). Thus, the use of composite measures might make the interpretation of results difficult. When the relationship between perceived fluency and utterance fluency is taken into account, fluency perceived by both native speakers and L2 learners is mostly linked to speed and breakdown fluency measures such as speech rate and pause ratio (Bosker et al., 2013; Derwing et al., 2004; Kormos & Dénes, 2004; Magne et al., 2019; Préfontaine et al., 2016; Rossiter, 2009; Saito et al., 2018). A recent meta-analysis of correlational studies further revealed that composite measures had the strongest relationships to perceived fluency (Suzuki et al., 2021). This is probably unsurprising because composite measures necessarily represent multiple dimensions of utterance fluency and are therefore likely to be more strongly associated with perceived fluency ratings than a measure that only represents one dimension.

Although the existing literature has demonstrated that certain utterance fluency measures can affect perceived fluency, much remains unknown about how both utterance and perceived fluency might be affected by pronunciation instruction. This is a particularly intriguing question because on the one hand, guided attention to pronunciation features might hinder fluency by directing learners' attention to detailed forms, but on the other hand, procedural knowledge of word linking and utterance chunking might improve fluency. In Derwing et al. (1998), when learners' narrative recordings were judged by six experienced ESL teachers based on a 9-point scale for fluency (from "NS-like fluency" to "extremely dysfluent"), it was found that only the group receiving suprasegmental instruction demonstrated gains in fluency. According to the authors, this is because suprasegmental knowledge can naturally be transferred to a spontaneous context compared to an awareness of segmental features. Another study that investigated the L2 fluency development of ESL learners in Canada using a narrative task was Rossiter (2009); however, unlike the segmental and suprasegmental instruction provided in Derwing et al. (1998), participants in Rossiter's study received 10 weeks of general English instruction (i.e., a communicative curriculum focused on four skills development). The findings from Rossiter's study on L2 fluency did not indicate any significant changes in perceived fluency. Nevertheless, Rossiter also analyzed the utterance fluency of the speech samples and asked listeners to comment on their ratings. Rossiter reported that perceived fluency ratings were correlated with some of the utterance fluency characteristics, but that some listener explanations attributed perceived fluency ratings to non-temporal features. These findings suggest that when investigating fluency, having both perceived and utterance fluency measures are useful. Therefore, the current study assessed learner fluency using both listener judgements (i.e., perceived fluency) and temporal measures of speech rate (i.e., speed fluency), filled and unfilled pauses (i.e., breakdown fluency), and repetitions, reformulations, replacements, and false starts (i.e., repair fluency).

As a final point, it seems that very few systematic investigations into the relationship between fluency and intelligibility have been conducted. A better understanding of how they are related can help justify clustering them (in addition to comprehensibility and accentedness) as a single construct of global L2 pronunciation proficiency (Saito & Plonsky, 2019). It can also help explain both previous and future research findings on global speech measures and benefit instruction in terms of resource allocation. For instance, Derwing and Munro (1997) found that only two out of 26 listeners' intelligibility scores significantly correlated with speech rate in their study of the relationship between accentedness, intelligibility, and comprehensibility, suggesting a weak relationship between intelligibility and fluency. Thomson (2015, p. 217) also concluded that "fluency is…apparently least related to intelligibility," but admitted that the evidence is limited. Thus, the current study explored the relationship between fluency and intelligibility. From the reviewed literature, three research questions were formulated:

1. What effects does L2 pronunciation instruction on segmental vs. suprasegmental features have on learner speech intelligibility and perceived as well as utterance fluency?

2. What relationship, if any, exists between L2 speech intelligibility and fluency?

3. To what extent can perceived fluency be predicted by utterance fluency (speed, breakdown, and repair) measures?

Method

Learners

Forty-five young adult EFL learners (L1=Vietnamese) were recruited from General English courses at an English language center in Vietnam to participate in a pronunciation course. Their proficiency was elementary to low intermediate, gauged by their completion of Level Two in the 4-level General English Program at this private institution (equivalent to Level A2 of the Common European Framework of Reference for Languages). Before voluntarily registering for the courses and signing the consent form for participation in a research study, the learners were informed about the course goals and syllabus. In the end, only the data collected from 30 learners was included in the analysis based on pre-established exclusion criteria (e.g., missing multiple classes, not completing weekly homework assignments). The learners varied in age from 16 to 26 years. This is a typical age range at the center, where both students and working people can attend. The learners were randomly assigned to one of two experimental groups: one group received instruction on segmental features (n=15) and the other received instruction on suprasegmentals (n=15). There was no control group due to the limited number of learners and the possibility that all learners may want to receive instruction. To provide all learners with the opportunity for instruction, it was not deemed appropriate to include a control group.

Instructors

For professional development purposes, there were a total of three nonnative instructors at the center involved in teaching the pronunciation classes. One main instructor co-taught each class with another instructor. The main instructor had an M.Ed. in Principles and Methods in English Language Education, whereas the other two instructors obtained their Bachelor's degree in English Studies. All instructors had taken at least three courses in phonetics in addition to an introduction to linguistics course during their undergraduate studies and were thus familiar with segmental and suprasegmental features of the English language. They also had previous experience teaching stand-alone pronunciation courses.

Instruction

The pronunciation courses lasted for eight weeks. Both the segmental and suprasegmental groups had two, 2-hour class sessions per week. The instruction was explicit (rule-based), and classroom procedures were adapted from Celce-Murcia et al. (2010): first, the target features and pronunciation rules were described to the learners; second, the learners completed listening tasks; third, the learners completed reading aloud; finally, they did communicative tasks on selected topics. Although the time spent on each step varied from class to class, read-aloud and communicative tasks each took approximately 30-45 minutes every class for both the groups. Corrective feedback was provided mostly in terms of recasts. For instance, when a student mispronounced the voiceless dental fricative /θ/, the instructor repeated their utterance but with corrected pronunciation. When the error persisted or was common among students, the instructor offered a metalinguistic explanation by drawing attention to target oral movements or articulation. Contrasts between similar sounds such as /θ/ and /ð/ were also explicitly discussed.

While the segmental group's syllabus revolved around groups of vowels and consonants, the suprasegmental group's comprised sets of rules for determining stress within words and sentences, dividing a sentence into thought groups, and shifting intonation for different sentence purposes (see Appendix A for complete course syllabi). The textbook used for the segmental group was *Ship or Sheep* (Baker, 2006). It covers all English vowels and consonants with each unit presenting an individual sound together with examples of the sound in words, minimal pairs, sentences, and dialogues. The suprasegmental group worked with *Clear Speech* (Gilbert, 2012), which covers a series of lessons on word and sentence stress, rhythm, intonation, and thought groups. Thus, each group trained with a rather comprehensive list of segmental/suprasegmental features. This approach was taken because the sound systems of English and Vietnamese differ greatly, causing Vietnamese-speaking learners of English to struggle in a variety of English pronunciation areas (Cunningham, 2009a, 2009b; Ehrlich & Avery, 2013). Moreover, the textbooks were readily available at the center, had clear and complete presentations of the target features, and provided various practice exercises and audio files. These audios served as a model of pronunciation in addition to the instructors' use of English as the primary medium of instruction. After every class, the learners in both groups were encouraged to practice the target features at home for roughly 15 minutes, but whether they spent more or less time practicing was not controllable. Each week, they also recorded (as homework) and received feedback on a practice speech on a given topic. Although the in-class practice time and the at-home practice topics were similar for both groups, the pronunciation foci evaluated were different. Table 1 summarizes the type of instruction for each group.

Table 1. Types of Instruction.

	Segmental group	**Suprasegmental group**
Hours of instruction	32 hours	32 hours
Instructional approach	Rule-based	Rule-based
Corrective feedback	Recast, metalinguistic explanation	Recast, metalinguistic explanation
Instructional focus	Short vowels, long vowels, diphthongs, consonants	Vowel rules, strong syllables, weak syllables, linking sounds
Textbook	*Ship or Sheep* (Baker, 2006)	*Clear Speech* (Gilbert, 2012)

Speech Collection

A pretest and posttest of the same design and procedure were administered to collect the learners' speech samples during the first and last weeks of class. According to Lee et al. (2015) and Thomson and Derwing (2015), the results of assessment in the form of controlled tasks (e.g., reading aloud a list of words or sentences) may not be indicative of learners' abilities in more communicative contexts; therefore, a less controlled instrument (i.e., monologues) was employed in the current study to increase authenticity and validity. Compared to picture narratives, monologues also offer more freedom as the narratives can impose unavoidable constraints on lexical items, structures, and content (Derwing et al., 2004). The same set of questions was used to elicit the speech of both groups during the pretest and posttest so that the speech could be more easily compared. See Chau (2021) for the recording instructions. Wh-questions were chosen because of their ability to trigger meaningful responses from the learners.

The pretest and posttest sessions were conducted one-on-one with one of the authors, who chatted with the learners in Vietnamese for a few minutes to make them feel at ease. The content of this initial brief chat (how the learners' day had been) was not related to the theme of the questions. The learners were then provided an English sheet of instructions along with further explanations in Vietnamese and given a couple of minutes to read through the questions, after which their responses to the questions were audio recorded, resulting in 30 pretest and 30 posttest recordings. The average length of the learners' pretest recordings was 123 seconds (*SD*=45.3) and the average length of their posttest recordings was 126 seconds (*SD*=32.9).

Stimulus Preparation

The pretest and posttest recordings were used to prepare stimuli for both the perceived fluency rating task and the intelligibility transcription task. For the stimuli used in the rating of fluency, Praat (Boersma & Weenink, 2017) was used to extract excerpts of around 30 seconds from each of the 60 recordings, beginning after the learners introduced themselves. These excerpts, from which all initial silent and filled pauses (e.g., *ah*, *um*) were removed, were cut at clausal boundaries. The actual mean length of the selected excerpts was 27.8 seconds (*SD*=2.71). For intelligibility transcription, a segment was further extracted from the beginning of each of the fluency excerpts. Similar to Munro and Derwing (1995), the segments selected were complete clauses of short duration (*M*=6.55 seconds, *SD*=1.99) appropriate for transcription, which ranged from 6 to 12 words (*M*=9.50, *SD*=1.94). Both sets of fluency and intelligibility stimuli were then normalized using Praat with a new absolute peak of 0.99, so the listeners would not have to adjust the volume while moving from one item to the next.

Listeners

Eleven English native speakers from a large public university in the USA were recruited to rate the fluency of the learners' speech and transcribe stimuli for the intelligibility task. They were undergraduate students enrolled in linguistics courses. Although all had been exposed to at least one particular group of L2 English learners (e.g., L1 Spanish or Chinese learners of English) and rated their familiarity with foreign-accented English 7.55 (*SD*=1.29) on a scale of 1 (not at all familiar) to 9 (very familiar), their level of familiarity with Vietnamese-accented English was low (*M*=3.73, *SD*=2.61).

Speech Assessment

There were two assessment sessions held in a quiet language lab on two different days, one for the fluency rating task and the other for the intelligibility transcription task. Both the fluency rating task and the intelligibility transcription task were presented to listeners in two separate experiments via Qualtrics. In both sessions, each listener was assigned to a computer with Internet connectivity and audio equipment and familiarized themselves with the procedures through three practice items. They were encouraged to complete the tasks at their own speed with short breaks in between. During the first session on the first day (the fluency rating task), the listeners heard each of the 60 fluency stimuli presented in a random order and were asked to listen to each stimulus once before rating it in terms of fluency using a 1000-point scale (Saito, Trofimovich, & Isaacs, 2016) with the leftmost end of the scale being extremely disfluent and the rightmost end being extremely fluent (see Figure 1). After completing the fluency rating, the listeners completed a short language background questionnaire in which they indicated information such as their language learning experiences and familiarity with L2 English speech. They also judged whether they understood the concept of fluency well (*M*=8.00, *SD*=.89) and their difficulty in completing the fluency rating task (*M*=5.64, *SD*=.89) on 9-point scales with 1 being "I did not understand this concept at all/very difficult" and 9 being “I understood this concept well/very easy.” The entire session lasted approximately one hour for each listener.

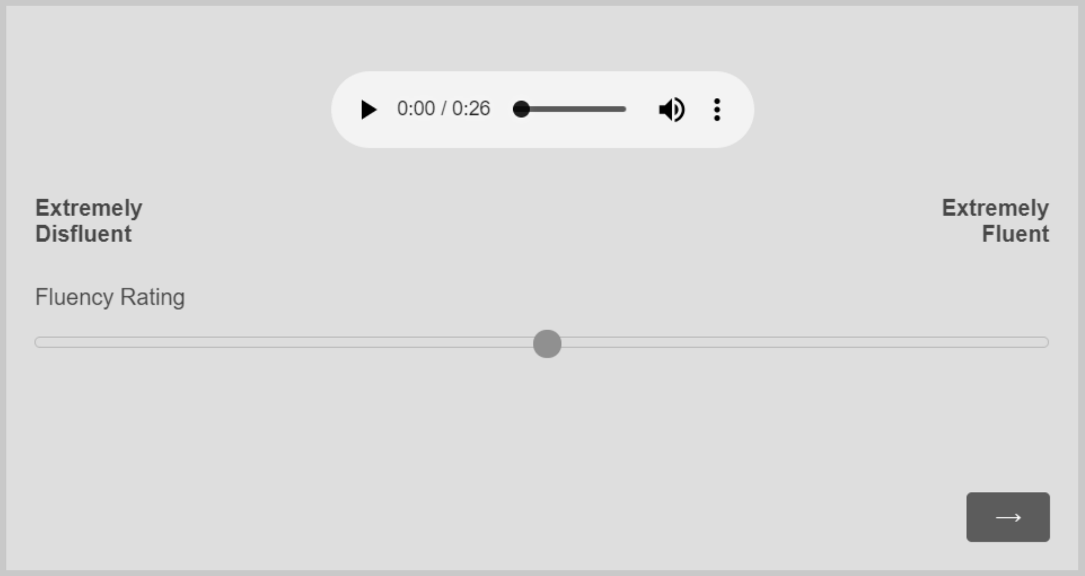


Figure 1. Sample Item from the Fluency Rating Task with Rating Scale.

The second session (the intelligibility transcription task) was held 1-2 days after the first session. During this session, the same listeners heard the intelligibility stimuli in a randomized order and were asked to write out exactly what they heard in standard orthography. Listeners were instructed to listen to the stimuli only one time. They found transcribing the intelligibility speech samples relatively easy (*M*=7.00, *SD*=.77) on a 9-point scale with 1 being “very difficult” and 9 being “very easy.” It took each listener approximately 30 minutes to complete the entire session.

Speech Coding and Analysis

Perceived fluency ratings and intelligibility transcriptions. Given the coding scheme adapted from Munro and Derwing (1995), the listeners’ transcriptions were compared to the original transcriptions for exact word match, omission, addition, and substitution of a content or function word, and regularization of an ungrammatical word. The intelligibility coding scheme is available in Chau (2021). The first two authors separately compared one random listener’s transcription with the corresponding original transcription and coded the difference for training before comparing and coding another three pairs of transcriptions for interrater reliability, which was high, κ=.93 (Cohen’s kappa, Plonsky & Derrick, 2016). Perceived fluency ratings and intelligibility transcriptions were analyzed for rater consistency. The listeners were generally consistent in their intelligibility transcriptions and fluency judgments with high reliability estimates (Cronbach’s alpha) of .93 and .88, respectively. Therefore, mean intelligibility and perceived fluency scores were calculated for each learner by averaging across all listeners’ exact word match percentages and ratings. Given the scores, two separate mixed/between repeated measures ANOVAs were conducted in SPSS with Time (two levels: pretest, posttest) as a within-subject factor and Group (two levels: segmental, suprasegmental) as a between-subject factor. The numerical results of Shapiro-Wilk normality tests suggested that all but the pretest suprasegmental intelligibility data subset (*p*=.02) and pretest segmental fluency (*p*=.03) were normally distributed. However, examination of histograms revealed that only the pretest suprasegmental and the posttest segmental intelligibility data were moderately negatively skewed. Transforming these variables did not result in improved distributions. As ANOVA is relatively robust to violations of normality and our interpretation of results focuses on effect sizes and confidence intervals, we report the ANOVA results while remaining cautious in our interpretation. An examination of box plots demonstrated similar variances among the groups.

Utterance fluency. In addition to listener fluency ratings and intelligibility transcriptions, the stimuli were also coded for temporal measures of fluency for an analysis of utterance fluency. First of all, the first and second authors individually coded six transcribed fluency stimuli (chosen at random). After discussing any discrepancies, another 12 stimuli were coded. Because there was good agreement, κ=.84, the first author continued to code the remaining stimuli. Each of the transcriptions were coded for filled pauses (e.g., *um*), silent pauses (silences of 250ms or longer, De Jong & Wempe, 2009), and instances of repair such as false starts, replacements, repetitions, and reformulations following a coding protocol. The utterance fluency coding scheme is available in Chau (2021).

Utterance fluency scores were calculated for each learner. Speed fluency, operationalized as speech rate, was measured by dividing the total number of pruned syllables, excluding repairs and filled pauses, by the total amount of elapsed time in seconds and multiplying by 60. Similarly, breakdown fluency was operationalized by dividing the total number of both filled and unfilled pauses by the total seconds of speaking time and multiplying by 60. Repairs were calculated by counting the total number of repetitions, reformulations or self-corrections, replacements, and false starts per minute. The number of syllables, pauses, and repairs per minute, representing utterance fluency, were then subjected to three separate repeated measures ANOVAs in SPSS parallel to the ones used for intelligibility and perceived fluency data. The results of Shapiro-Wilk normality tests and inspections of histograms suggested that several of the repair fluency data subsets were positively skewed. Transforming these variables did not result in improved distributions. As with the intelligibility results, we report the ANOVA results while remaining cautious in our interpretation. An examination of box plots confirmed that variances were similar among the groups.

Interpretation of statistical tests. For all omnibus statistical tests, an alpha level of .05 was used. In the case of post-hoc comparisons, a Bonferroni correction was applied. To answer the first research question, effect sizes (Cohen's *d*) along with 95% CIs were included to quantify the practical significance of segmental vs. suprasegmental instruction on speech intelligibility and fluency. Effect sizes are interpreted based on Lee et al.'s (2015) meta-analysis, in which a mean within-group effect size of .89 (95% CI [0.85, 0.94]) was reported. For the second and third research questions, which examined if there were any relationships between the learners' speech intelligibility and fluency as well as between their utterance and perceived fluency, scatterplots and multiple regression were employed, respectively.

Additionally, in an effort to attend to the inference crisis of the social and behavioral sciences (Norouzian et al., 2019; Rouder et al., 2016), five Bayesian repeated measures ANOVAs were run using JASP (JASP Team, 2020) to complement conventional null hypothesis testing that relies on *p*-values. Bayesian hypothesis testing specifies the alternative hypotheses, obtains a comparative measure or Bayes factor, and interprets the Bayes factor (Norouzian et al., 2019). While Bayes factors are direct measures of the evidence against the null/alternative hypothesis, *p*-values are just indirect measures of the evidence against the null hypothesis because it is estimated under the assumption that the null hypothesis is true (Held & Ott, 2018). In other words, *p*-values provide limited information and thus can be augmented or substituted with Bayes factors to quantify the relative evidence for both the null and alternative hypotheses. The Bayes factors obtained in this study are interpreted based on Norouzian et al.'s (2019, p. 252) Bayes Factor Classificatory Scale, which range from BF<.01 (decisive evidence for the null hypothesis) to BF>100 (decisive evidence for the alternative hypothesis).

Results

The Effects of Instruction on Intelligibility and Fluency

The following subsections report the results of the quantitative analyses related to the first research question, which investigated the effects of segmental and suprasegmental pronunciation instruction on intelligibility, perceived fluency, and utterance fluency.

Intelligibility. Results from the ANOVA analysis on intelligibility transcriptions yielded no main effect of Time, $F(1,28)=1.41$, $p=.25$, $\eta_p^2=.05$, indicating that the mean percentages of exact word match did not vary significantly between the pretest ($M=83.4$, $SD=14.0$) and posttest ($M=79.9$, $SD=13.6$). The interaction between Time and Group was also nonsignificant, $F(1,28)=.19$, $p=.67$, $\eta_p^2=.007$. There was no main effect of Group, $F(1,28)=.24$, $p=.63$, $\eta_p^2=.008$, indicating that the mean word match percentage was not significantly different for the segmental group ($M=82.7$, $SD=13.1$) compared to the suprasegmental group ($M=80.7$, $SD=14.7$). To confirm these results, a Bayesian repeated measures ANOVA was conducted. Strong to decisive evidence for the null hypothesis (i.e., no difference in intelligibility) was found for all Time (BF10=.05), Group (BF10<.01), and the Time x Group interaction (BF10=.02) (Norouzian et al., 2019).

Table 2 reports the frequencies of seven types of transcription errors, of which omission and substitution of content words accounted for nearly half of the frequencies (44.5%).

Table 2. Frequencies of Transcription Error Types.

	Omission		Addition		Substitution		Reformulation
	Content word	Function word	Content word	Function word	Content word	Function word	N/A
Freq	270	194	17	57	218	93	249
%	24.6	17.7	1.5	5.2	19.9	8.5	22.7

However, the generally high exact word match rate (about 80%) across time and group indicated that the speech samples had relatively high intelligibility. More specifically, the distribution of intelligibility scores is highly skewed, with 71% of the distribution including scores from 75 to 100 (see Figure 2).

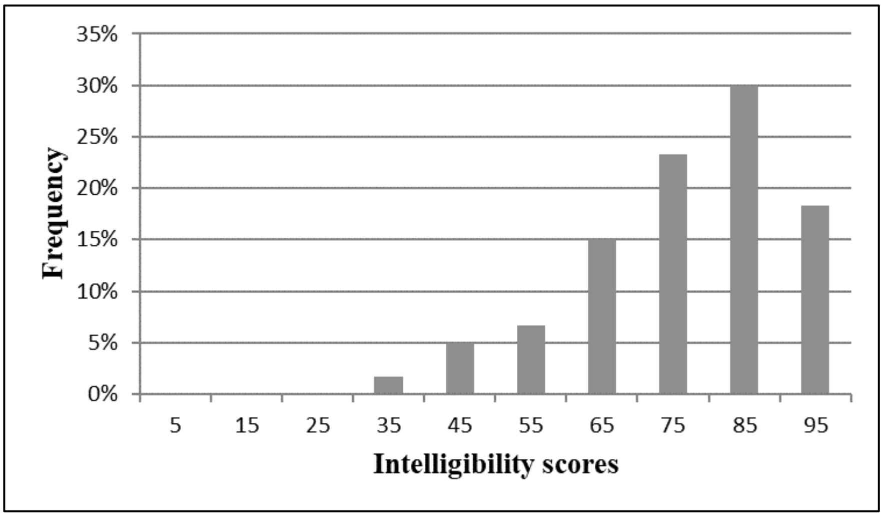


Figure 2. Distribution of Intelligibility Scores Across Time and Group.

Perceived fluency. The results for perceived fluency yielded no main effect for Time, $F(1,28)=.75$, $p=.39$, $\eta_p^2=.03$, such that the average rating was not significantly higher for the posttest (*M*=505.2, *SD*=126.0) than for the pretest (*M*=484.1, *SD*=136.7). There was no main effect of Group, *F*(1,28)=.44, *p*=.51, η_p^2=.02. The mean rating was not significantly different for the segmental group (*M*=508.5, *SD*=120.5) compared to the suprasegmental group (*M*=480.8, *SD*=140.9). The interaction effect was also nonsignificant, *F*(1,28)=1.75, *p*=.20, η_p^2=.06. A Bayesian repeated measures ANOVA confirmed these results with weak or anecdotal to decisive evidence for no effect of Time (BF10=.05), Group (BF10<.01), and Time x Group interaction (BF10=.47) (Norouzian et al., 2019). This means no difference in perceived fluency was observed over time for the segmental group, the suprasegmental group, and between the groups.

Utterance fluency. The ANOVA for the number of pruned syllables per minute revealed no significant effects for Time, *F*(1,28)=.97, *p*=.33, η_p^2=.03, or Group, *F*(1,28)=.19, *p*=.67, η_p^2=.007, but a significant Time x Group interaction was found, *F*(1,28)=5.22, *p*=.03, η_p^2=.16 (see Table 3 for the means and *SD*s of utterance fluency measures by time and group). Such results indicated that the groups had significantly different patterns of change over time, but there were generally no significant changes from pretest to posttest or significant differences in the average number of syllables per minute between them. To determine the groups' change patterns, two post-hoc Bonferroni-adjusted paired samples *t*-tests were performed to compare differences in the number of syllables per minute within each group across time. The results did not suggest a significant change for the segmental group in terms of speed fluency, *t*(14)=.98, *p*=.34, *d*=-0.26, 95% CI [-0.97, 0.46]. Similarly, despite a *p* value of .05, a negligible effect size whose CIs passed through zero (*d*=.54, 95% CI [-0.21, 1.25]) indicated that the increase in speech rate from the pretest to the posttest of the suprasegmental group was not meaningful.

The ANOVA for the number of pauses per minute revealed no significant effects for Time, *F*(1,28)=1.69, *p*=.20, η_p^2=.06, Group, *F*(1,28)=.96, *p*=.34, η_p^2=.03, or Time x Group interaction *F*(1,28)=.70, *p*=.41, η_p^2=.03.

The ANOVA for the number of repairs per minute revealed no significant effects for Time, *F*(1,28)=.93, *p*=.76, η_p^2=.003, and no significant Time x Group interaction, *F*(1,28)=.23, *p*=.64, η_p^2=.008. Once again, no significant effects for Group were found, *F*(1,28)=1.46, *p*=.24, η_p^2=.05.

Table 3. Descriptive Statistics for Utterance Fluency Measures.

	Test	*n*	Mean	*SD*	*n*	Mean	*SD*
		Time x Group: segmental			Time x Group: suprasegmental		
Speed fluency (pruned syllables /minute)	Pretest	15	92.2	21.4	15	85.5	26.9
	Posttest	15	86.5	22.1	15	99.9	26.7
Breakdown fluency (pauses/minute)	Pretest	15	40.5	8.62	15	45.3	13.6
	Posttest	15	39.7	8.96	15	41.4	8.82
Repair fluency (repairs/minute)	Pretest	15	5.42	3.28	15	6.95	6.36
	Posttest	15	5.25	4.52	15	7.72	6.26

Results from Bayesian ANOVAs are confirmatory (see Table 4). They provided substantial to decisive evidence for the lack of improvement in utterance fluency and difference between the segmental and suprasegmental groups (Norouzian et al., 2019).

Table 4. Bayesian ANOVAs of Utterance Fluency.

	Time	Group	Time x Group
Speed fluency	BF10=.26	BF10=.81	BF10=.21
Breakdown fluency	BF10=.02	BF10<.01	BF10=.18
Repair fluency	BF10=.28	BF10=.27	BF10=.13

Relationship Between Intelligibility and Fluency

The second research question investigated the relationships of intelligibility to perceived fluency and utterance fluency. Scatterplots (see Appendix B) showed that data points were not linear, nor was there an obvious upward or downward trend in the data. Most of the points are concentrated in the upper part of the charts. This distributional pattern corroborates previous observations that the participants scored relatively high in intelligibility. Since the assumption of linearity was not met, it was inappropriate to test for linear relationships by performing correlations (Larson-Hall, 2016). Nevertheless, visual inspection of the plots appears to indicate that intelligibility scores did not vary according to fluency scores.

Relationship Between Utterance Fluency and Perceived Fluency

Standard multiple regression analyses were performed to predict perceived fluency ratings from utterance fluency temporal measures representing speed, breakdown, and repair fluency. Only the speed fluency (i.e., speech rate) model significantly predicted perceived fluency, $F(1, 58)=112.0, p < .001$, adj. $R^2=.65$. Further hierarchical multiple regression analyses were run to determine if the addition of breakdown and repair fluency measures improved the prediction of perceived fluency over and above speed fluency or speech rate alone. However, such additions (models 4, 5 and 6) did not lead to a statistically significant increase in R^2. See Appendix C for details on each regression model.

Discussion

The main objective of this study was to better understand the effects of both segmental and suprasegmental approaches to pronunciation instruction on the intelligibility and fluency of spontaneous speech elicited from EFL learners in Vietnam. A secondary goal was to explore the relationships between intelligibility and fluency as well as between perceived and utterance fluency. Overall, the results indicated neutral effects of both types of instruction on intelligibility and fluency, with the only effect trending toward practical significance being an increase in speed fluency for the suprasegmental group. No clear relationship was found between intelligibility and fluency, and while speech rate was a robust predictor of perceived fluency, breakdown and repair fluency were not significant predictors.

We first address the results related to research question 1 and the finding that the pronunciation instruction provided in the current study did not lead to significant gains in intelligibility or perceived fluency for these learners. At first glance, these null results might seem surprising given that previous systematic reviews have indicated that pronunciation instruction is generally effective (Lee et al., 2015; Saito, 2012; Thomson & Derwing, 2015). However, in the current study, the learners' speech was evaluated using the global measures of intelligibility and fluency. While it is the case that a majority of previous pronunciation instruction studies have reported significant improvement in pronunciation, their heavy focus on evaluating discrete features (e.g., English /ɹ/ or Spanish /d/) using controlled tasks may mean that the same effects are less likely to generalize to global measures of speech (Thomson & Derwing, 2015). These findings are in line with Saito and Plonsky (2019) whose meta-analysis indicated a general lack of significant effects of pronunciation instruction on global speech dimensions such as comprehensibility. Saito and Plonsky (2019) proposed a model of L2 pronunciation proficiency which incorporates two important concepts relevant to the findings in the current study. The first is that in their proposed model, instruction does not directly improve global L2 pronunciation, but rather (if effective) instruction improves specific segmental and suprasegmental features, which in turn feed into potential improvements in global dimensions. A second important consideration is that global dimensions of L2 speech are impacted by more than only pronunciation features such that other variables (e.g., lexicogrammatical features, rater familiarity with L2 speech) also feed into evaluations of the global dimensions. In this way, the lack of significant improvement found in the current study is not likely an indication that the learners' speech stayed the same, but rather that any changes that did occur were too small to surface in the global dimensions.

Another potential explanation regarding the null findings for research question 1 relates to important considerations regarding the type of instruction and practice the learners received in connection to the spontaneous speaking task used as an outcome measure. As described in the method section, learners first received explicit instruction on the target pronunciation features, followed by listening practice and reading aloud, and then finished with communicative practice. Nevertheless, there was more use of drills compared to communicative and authentic language tasks as the third step – reading aloud proved to be challenging and thus time-consuming for the learners. It was probably because the learners were trying to balance producing speech smoothly with accurately applying their newly gained knowledge of English segmentals/suprasegmentals. Drilling was also present throughout the lesson. For example, the learners were often asked to repeat after the instructor during the first step, when new sounds or pronunciation features were introduced. Given that the majority of learners' practice was controlled in combination with the fact that pretest and posttest outcome measures were comprised of a spontaneous task, it is perhaps not surprising that significant gains in intelligibility and fluency were not found. In comparison to Derwing et al. (1998) who did find significant improvement in perceived fluency for the suprasegmental group in their study, context might have played an additional role: Learners in their study were in an ESL context whereas learners in the current study were in the Vietnamese EFL context. Perhaps additional access to speaking opportunities outside of class played a role. This explanation aligns with findings from previous systematic reviews of the pronunciation instruction literature that have reported larger effects in second vs. foreign language contexts (Lee et al., 2015). These findings raise an interesting question about the relative impact of pronunciation instruction compared to that of contextual variables, which should be examined in more detail in future studies.

Regarding the results for intelligibility specifically, a closer look at the relatively high transcription scores in the pretest (83.4%) perhaps suggests that the lack of significant improvement in intelligibility might be due to a ceiling effect. Although the proficiency level of the learners in current study ranged from elementary to low intermediate, their overall intelligibility scores were quite high across all times and groups (*M*=81.7, *SD*=13.8), and the listeners found it relatively easy to transcribe learners' speech despite their low familiarity with Vietnamese-accented English. Perhaps surprisingly although encouragingly, the intelligibility scores of the participants in the current study are quite similar to the scores of the advanced speakers reported in Munro and Derwing (1995, pp. 83–84). In other words, the low-level language learners in the current study appear to be as intelligible as the more proficient learners in the Munro and Derwing study. These results potentially indicate a weak relationship between proficiency and intelligibility. Perhaps, proficiency is more associated with comprehensibility (i.e., how easy or difficult it is to understand a speaker) rather than intelligibility. It is also possible that the relatively consistent and high scores for intelligibility across both studies are the result of how intelligibility was operationalized. The short duration of the speech segments used as stimuli (*M*=10.7 words for Munro and Derwing and *M*=9.50 words for the current study) and the similarity in linguistic content found across different speakers might have made transcribing the speech less challenging. Although no single approach to measuring intelligibility is without its disadvantages (see e.g., Kang et al., 2018), such a shortcoming could be addressed in future research with the use of multiple measures of intelligibility.

Turning to the results of perceived fluency, the stability in ratings from pretest to posttest may also reflect the raters' attitudes toward the fluency rating task. The task ran on Saito, Trofimovich, and Isaacs's (2016) 1000-point scale; however, many raters, when asked about their impressions, commented that they found it difficult to judge the participants subjectively. When completing the intelligibility task, they simply had to transcribe what they heard, which removed them from a judgmental position; whereas, when rating fluency they had to play a more decisive role. The fact that the average ratings in the pretest (*M*=484.1) and posttest (*M*=505.2) were close to the middle of the scale (with *SD*s just over 100) provides some indication that the raters did not allow much variability in their judgements. If rater attitudes had any impact on scale use, researchers in future studies could explicitly encourage raters to use the entire range of the scale by emphasizing that raters should feel comfortable providing their honest judgements.

In terms of utterance fluency, neither of the groups was able to speak faster and pause or repair less after the instruction. One possible explanation for this finding is that the segmental group was instructed to focus on the articulation of individual vowels and consonants, and this new knowledge of discrete sounds might have made learners more aware of and more likely to reformulate any incorrect pronunciations. Such an awareness of producing accurate target sounds was less likely to be present in the mind of those in the suprasegmental group who received instruction on stress, rhythm, and intonation. Nevertheless, although their speed fluency trended toward improvement, it seems that eight weeks of instruction (4 hours a week), a heavy reliance on decontextualized drills, and limited exposure to authentic, fast-paced English were not enough to improve the suprasegmental learners' fluency significantly. In parallel to the findings for perceived fluency, language learning context might have played a role: Being in a foreign language environment might have limited the opportunities for outside-the-classroom English interactions.

The second research question focused on the extent to which the global speech dimensions of intelligibility and fluency are related. It was not possible to statistically test the relationship between intelligibility and fluency because, as discussed earlier, the assumption of linearity was not met in the data, but scatterplots (see Appendix B) provided some evidence that intelligibility scores did not vary according to fluency as no clear trend was visible in data. Very few studies have directly explored the relationship between intelligibility and fluency, and based on the limited, indirect evidence available, Thomson (2015) hypothesized that their relationship may be weak. The findings in the current study appear to support this claim. Empirical evidence has instead indicated that fluency is most related to comprehensibility (Thomson, 2015). In practical terms, it seems that one tentative conclusion is that fluency might not stop people from understanding each other (i.e., intelligibility), but it does seem to affect the effort needed to understand (i.e., comprehensibility). Nevertheless, these conclusions are tentative and thus more robust evidence would need to be provided to confirm or refute these claims.

Finally, the third research question explored the relationship between perceived and utterance fluency, and as expected, speech rate was found to strongly predict perceived fluency. Although this finding aligns with previous research (Bosker et al., 2013; De Jong et al., 2013;

Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009), it should be interpreted with caution. Since speech rate as measured in the current study is a composite measure, its strong association with perceived fluency scores can be attributed to multiple aspects of utterance not just speed alone. The finding suggests that TESOL teachers who want to devote more time to fluency might need to focus on multiple aspects of fluency.

With an eye to future research, one important question to explore is the impacts of length of instruction and different combinations of instruction types in the classroom to determine the optimal approach to pronunciation instruction. The pronunciation instruction in the current study was influenced by the resources available at the language institute. Both the instructors and institution management agreed that it was best for the learners to take advantage of the textbooks available and their exercises. Nevertheless, the growth of the field has motivated researchers and practitioners to stake out new territory. A modern view of pronunciation instruction may take into consideration other empirically supported methods beyond those found in textbook exercises such as those deriving from computer technology (e.g., McCrocklin, 2016) and drama and theater (e.g., Galante & Thomson, 2017; Gill, 2013). Future studies investigating the effectiveness of pronunciation instruction could compare the effects of these methods on different speech measures in different measurement tasks.

Admittedly, as with any other study, the current study has limitations that should be acknowledged. One of them is the lack of a control group that did not receive any pronunciation instruction due to logistic and fairness issues. In the reviewed literature, pronunciation instruction has been shown to be effective to some extent, and comparing the effects of two different types of pronunciation instruction using two experimental groups was the design chosen for this study. This is an alternative approach to comparing the effects of an intervention to no instruction at all. In this way, while we were able to explore similarities and differences between the two instruction types, we are not able to comment on the effects of instruction versus lack of it. For this reason, future research should strongly consider including a control group to be able to address both questions. Also, the small sample size of this study might have made it difficult to find statistically significant differences between pre and posttest fluency and intelligibility scores. Thus, a series of Bayesian ANOVAs were run to obtain evidence for the absence of difference and enable more reliable inferences. Researchers can take similar approaches when encountering the same issue (see McNeish, 2016; Norouzian et al., 2019; van de Schoot & Miočević, 2020). Another potential limitation was the use of the same set of eliciting questions to assess learners' pronunciation before and after intervention. While the 2-month intervention period between the pretest and posttest was expected to mitigate any practice effects, prospective researchers following a similar design may consider dividing the set of questions into two subsets and counterbalancing them across pre and posttests. Finally, it would be interesting to investigate the relationship between teacher expertise/training and learner improvement as the instructors in this study were experienced in language teaching but had not been specifically trained on pronunciation instruction.

As noted at the outset, the ultimate goal of pronunciation instruction should be helping learners achieve intelligible and comprehensible speech. Applying a pretest-posttest design, this study showed no significant effects of an 8-week course in segmental or suprasegmental instruction on learners' speech intelligibility and fluency. This study has provided a more thorough understanding of the effects of pronunciation instruction on two global measures of speech in an EFL context and the importance of speech rate to perceived fluency. Despite the null results reported in the current study, the findings do appear to indicate that lower proficiency L2 speakers can be intelligible even when the speaking task is spontaneous. Such a finding is encouraging.

About the Authors

Tuc Chau is a PhD candidate in Linguistics and Applied Language Studies at the University of South Florida. His research interests include second language writing and pronunciation.

Amanda Huensch is Assistant Professor in the Department of Linguistics at the University of Pittsburgh. Her research examines second language speech development in and outside of the classroom, including pronunciation attitudes of classroom learners and fluency development during study abroad.

Yen K. Hoang is an instructor at Gia Viet English Language Center.

Hiep T. Chau is a lecturer in the School of Foreign Languages at Can Tho University.

Acknowledgements

We would like to thank our participants. We would also like to thank the anonymous reviewers for taking time and effort to help us improve the article.

To cite this article

Chau, T., Huensch, A., Hoang, Y. K. & Chau, H. T. (2022). The effects of L2 pronunciation instruction on EFL learners' intelligibility and fluency in spontaneous speech. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 25*(4). https://tesl-ej.org/pdf/ej100/a7.pdf

References

Amini, M., & Birjandi, P. (2012). Gender bias in the Iranian high school EFL textbooks. Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning, 59*(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

Baker, A. (2006). *Ship or sheep? An intermediate pronunciation course* (3rd ed.). Cambridge University Press.

Boersma, P. & Weenink, D. (2017). *Praat: Doing phonetics by computer* (Version 6.0.36) [Computer software]. University of Amsterdam. http://www.fon.hum.uva.nl/praat

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

Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge University Press.

Chau, T. (2021). The effects of L2 pronunciation instruction on EFL learners' intelligibility and fluency in spontaneous speech [Data set]. Open Science Framework (OSF). https://osf.io/bjwc3/

Cunningham, U. (2009a). Phonetic correlates of unintelligibility in Vietnamese-accented English. In P. Branderud & H. Traunmüller (Eds.), *Proceedings of FONETIK 2009: The XXIIth Swedish phonetics conference* (pp. 108–111). Stockholm University. http://su.diva-portal.org/smash/get/diva2:273431/FULLTEXT01.pdf

Cunningham, U. (2009b). Quality, quantity and intelligibility of vowels in Vietnamese-accented English. In E. Waniek-Klimczak (Ed.), *Issues in accents of English II: Variability and norm* (pp. 3–22). Cambridge Scholars Publishing.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics, 34*(5), 893–916. https://doi.org/10.1017/S0142716412000069

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods, 41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Derwing, T. M., & Munro, M. J. (1997). Accent, comprehensibility and intelligibility: Evidence from four L1s. *Studies in Second Language Acquisition, 19*(1), 1–16. https://doi.org/10.1111/00238333.49.s1.8.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163–185. https://doi.org/10.1111/lang.12000

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins Publishing Company.

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning, 48*(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*(4), 655–679. https://doi.org/10.1111/j.1467-9922.2004.00282.x

Ehrlich, S., & Avery, P. (2013). *Teaching American English pronunciation-Oxford handbooks for language teachers*. Oxford University Press.

Galante, A., & Thomson, R. I. (2017). The effectiveness of drama as an instructional approach for the development of second language oral fluency, comprehensibility, and accentedness. *TESOL Quarterly, 51*(1), 115–142. https://doi.org/10.1002/tesq.290

Gilbert, J. B. (2012). *Clear speech: Pronunciation and listening comprehension in North American English* (4th ed.). Cambridge University Press.

Gill, C. (2013). Enhancing the English-language oral skills of international students through drama. *English Language Teaching, 6*(4), 29–41. https://doi.org/10.5539/elt.v6n4p29

Held, L., & Ott, M. (2018). On *p*-values and Bayes factors. *Annual Review of Statistics and Its Application, 5*, 393–419. https://doi.org/10.1146/annurev-statistics-031017-100307

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). John Benjamins Publishing Company.

Huensch, A., & Tracy-Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *The Modern Language Journal*, *101*(2), 275–293. https://doi.org./10.1111/modl.12395

JASP Team. (2020). *JASP* (Version 0.12.2) [Computer software]. University of Amsterdam. https://jasp-stats.org

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Jenkins, J. (2007). *English as a lingua franca: Attitude and identity*. Oxford University Press.

Jenkins, J., Cogo, A., & Dewey, M. (2011). Review of developments in research into English as a lingua franca. *Language Teaching, 44*(3), 281–315. https://doi.org/10.1017/S0261444811000115

Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning 68*(1), 115–146. https://doi.org/10.1111/lang.12270

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164. https://doi.org/10.1016/j.system.2004.01.001

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics 36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*(3), 369–377. https://doi.org/10.2307/3588485

Levis, J. M. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation, 6*(3), 310–328. https://doi.org/10.1075/jslp.20050.lev

Magne, V., Suzuki, S., Suzukida, Y., Ilkan, M., Tran, M. N., & Saito, K. (2019). Exploring the dynamic nature of second language listeners' perceived fluency: A mixed-methods approach. *TESOL Quarterly, 53*(4), 1139–1150. https://doi.org/10.1002/tesq.528

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System, 57*, 25–42. https://doi.org/10.1016/j.system.2015.12.013

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal, 23*(5), 750-773. https://doi.org/10.1080/10705511.2016.1186549

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.14671770.1995.tb00963.x.

Munro, M. J., & Derwing, T. M. (2015). Intelligibility in research and practice: Teaching priorities. In M. Reed & J. Levis (Eds.), *The handbook of English pronunciation* (pp. 375–396). John Wiley & Sons.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). *The mutual intelligibility of L2 speech. Studies in Second Language Acquisition, 28*(1), 111–131. https://doi.org/10.1017/S0272263106060049

Murphy, J. M. (2014). Intelligible, comprehensible, non-native models in ESL/EFL pronunciation teaching. *System, 42*, 258–269. https://doi.org/10.1016/j.system.2013.12.007

Norouzian, R., Miranda, M. D., & Plonsky, L. (2019). A Bayesian approach to measuring evidence in L2 research: An empirical investigation. *The Modern Language Journal, 103*(1), 248–261. https://doi.org/10.1111/modl.12543

Parlak, Ö. (2010). Does pronunciation instruction promote intelligibility and comprehensibility? *SPLIS News, 7*(2). https://www.tesol.org/news-landing-page/2011/11/03/splis-news-volume-7-2-(october-2010)

Plonsky, L., & Derrick, D. J. (2016). A meta-analysis of reliability coefficients in second language research. *The Modern Language Journal, 100*(2), 538–553. https://doi.org/10.1111/modl.12335

Préfontaine, Y., Kormos, J., & Johnson, D. E. (2016). How do utterance measures predict raters' perceptions of fluency in French as a second language? *Language Testing, 33*(1), 53–73. https://doi.org/10.1177/0265532215579530

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review 65*(3), 395–412. https://doi.org/10.3138/cmlr.65.3.395

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E. J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science, 8*(3), 520–547. https://doi.org/10.1111/tops.12214

Saito, K. (2012). Effects of instruction on L2 pronunciation development: A synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly, 46*(4), 842–854. https://doi.org/10.1002/tesq.67

Saito, K., Ilkan, M., Magne, V., Tran, M. N., & Suzuki, S. (2018). Acoustic characteristics and learner profiles of low-, mid-and high-level second language fluency. *Applied Psycholinguistics, 39*(3), 593–617. https://doi.org/10.1017/S0142716417000571

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning, 69*(3), 652–708. https://doi-org/10.1111/lang.12345

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*(2), 217–240. https://doi.org/10.1017/S0142716414000502

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Seidlhofer, B. (2013). *Understanding English as a lingua franca.* Oxford University Press.

Suzuki, S., Kormos, J., & Uchihara, T. (2021). The relationship between utterance and perceived fluency: A meta-analysis of correlational studies. *The Modern Language Journal, 105*(2), 435–463. https://doi.org./10.1111/modl.12706

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–276). John Benjamins Publishing Company.

Thomson, R. I. (2015). Fluency. In M. Reed & J. Levis (Eds.), *The handbook of English pronunciation* (pp. 209–226). John Wiley & Sons.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics, 36*(3), 326–344. https://doi.org/10.1093/applin/amu076

van de Schoot, R., & Miočević, M. (Eds.). (2020). *Small sample size solutions: A guide for applied researchers and practitioners.* Routledge.

Appendix A. Complete Syllabi for the Segmental and Suprasegmental Groups.

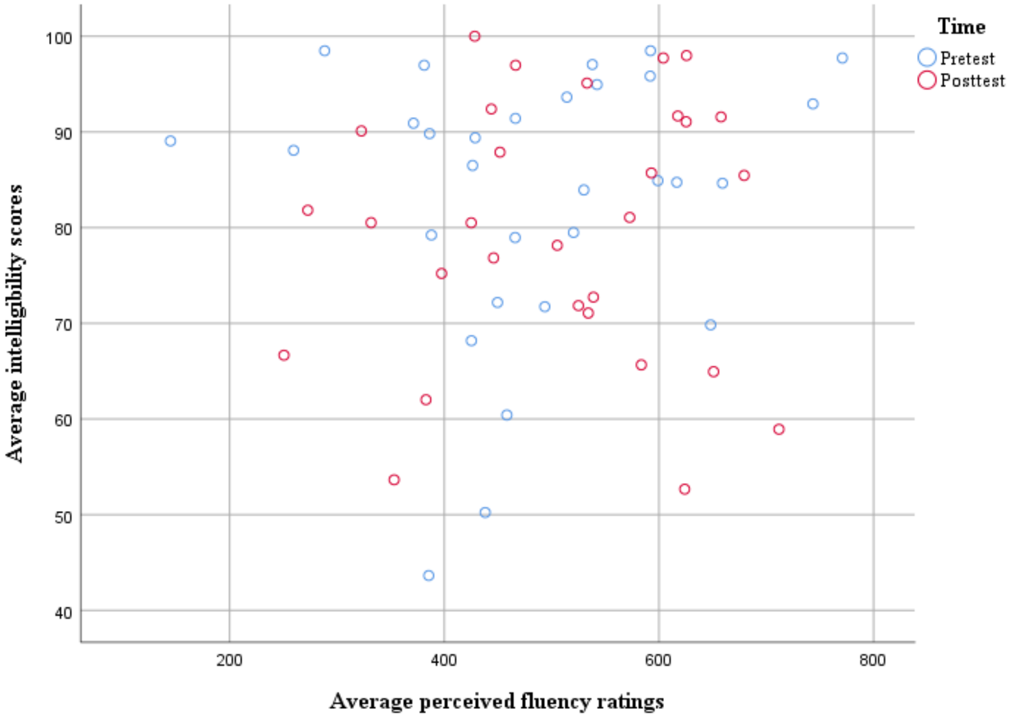
Segmental	Friday	Saturday
Week 1	Brief introduction on the course. Some exercises for training organs of speech. Vowels /i:, ɪ, e, æ, ʌ, ɑ:, ɒ, ɔ:/	Review and practice /i:, ɪ, e, æ, ʌ, ɑ:, ɒ, ɔ:/ Speaking topics for more practice: (1) Talk about your family; (2) Talk about the person you love most in your family
Week 2	Vowels (cont.) /ʊ, u:, ɛ:, ə/	Review /ʊ, u:, ɛ:, ə/ Diphthongs /eɪ, aɪ, ɔɪ/ Practice /ɒ, ɔ:, ʊ, u:, ɛ:, ə, eɪ, aɪ, ɔɪ/ Progress test 1 Speaking topics for more practice: (1) Describe how you spend your free time; (2) Talk about your favorite free time activity
Week 3	Review diphthongs /eɪ, aɪ, ɔɪ/ Diphthongs (cont.) /aʊ, əʊ, ɪə, eə/	Review and practice all diphthongs learnt /eɪ, aɪ, ɔɪ, aʊ, əʊ, ɪə, eə/ Progress test 2 Speaking practice topics: (1) Describe your house; (2) Talk about your dream house
Week 4	Consonants /p, b, t, d/	Review /p, b, t, d/ Consonants (cont.) /k, g/ Practice /p, b, t, d, k, g/ Progress test 3 Speaking practice topics: (1) Describe your typical day; (2) Talk about what you normally do on the weekend
Week 5	Consonants (cont.) /s, z, ʃ, ʒ/	Review /s, z, ʃ, ʒ/ Consonants (cont.) /tʃ, dʒ, f, v/ Progress test 3 Texts for more practice: 2 short stories from American Anecdotes – Elementary

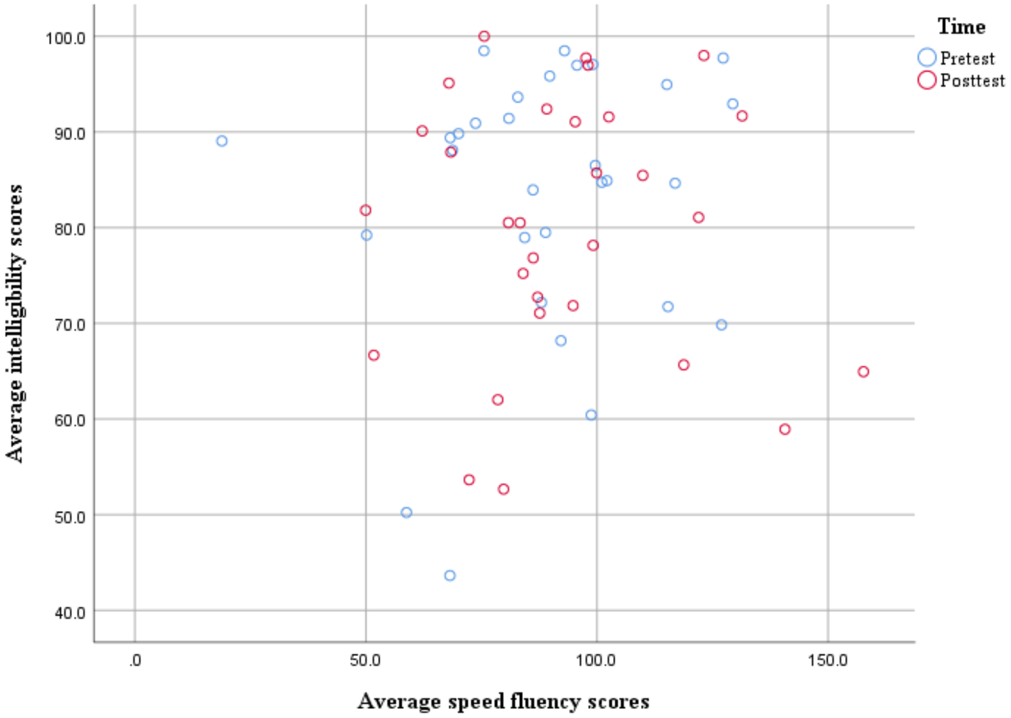
Week 6	Consonants (cont.) /w, j, h, m, n, ŋ/	Consonants (cont.) /Θ, ð, l, r/ Review and practice /w, j, h, m, n, ŋ, Θ, ð, l, r/ Progress test 4 Speaking topics for more practice: (1) Would you prefer to live in the countryside or in a big city. Explain why; (2) Where would you like to travel if you have a chance
Week 7	Review all vowels /i:, ɪ, e, æ, ʌ, ɑ:, ɒ, ɔ:, ʊ, u:, ɛ:, ə, eɪ, aɪ, ɔɪ, aʊ, əʊ, ɪə, eə/	Review all consonants /p, b, t, d, k, g, s, z, ʃ, ʒ, tʃ, dʒ, f, v, w, j, h, m, n, ŋ, Θ, ð, l, r/
Week 8	Wrapping up	

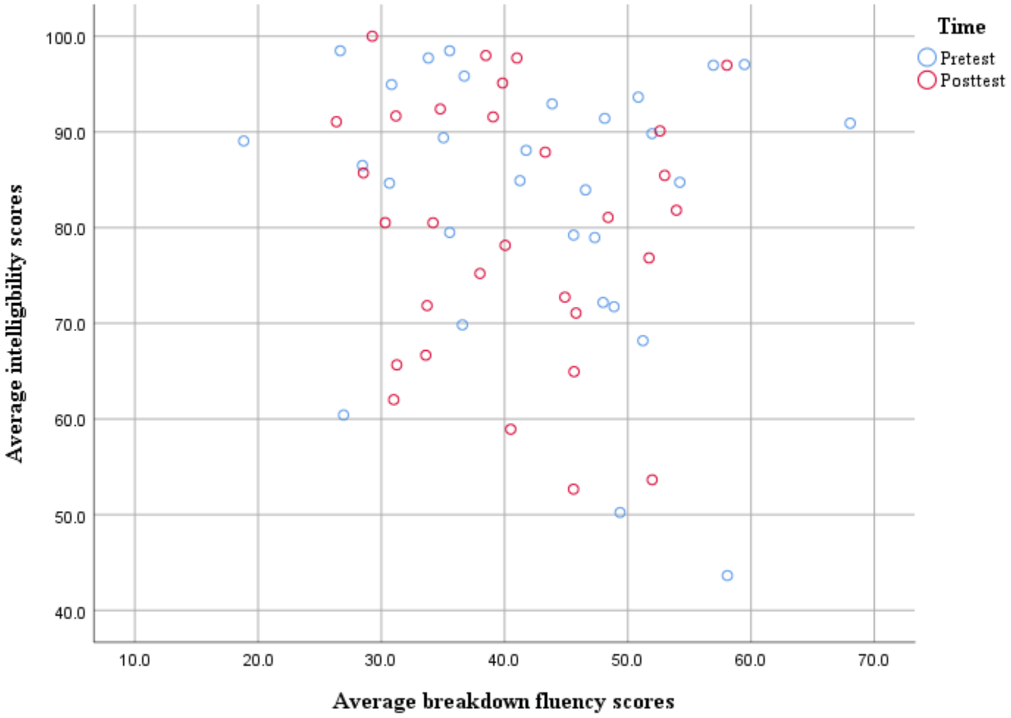
Suprasegmental	Saturday	Sunday
Week 1	Brief introduction on the course Some exercises for training organs of speech. Syllables	Review syllables Word stress Speaking topics for more practice: (1) Talk about your family; (2) Talk about the person you love most in your family
Week 2	Word stress (cont.): rules for stress within words	Practice on syllables and word stress Progress test 1 Speaking topics for more practice: (1) Describe how you spend your free time; (2) Talk about your favorite free time activity
Week 3	De-emphasizing: schwa Sentence stress: rules for choosing the focus word	Practice word and sentence stress Progress test 2 Speaking practice topics: (1) Describe your house; (2) Talk about your dream house
Week 4	De-emphasizing: contraction, reduction, silent letter h Sentence stress: disagreeing and correcting	Rhythm: music of English Practice sentence stress and rhythm Progress test 3 Speaking practice topics: (1) Describe your typical day; (2) Talk about what you normally do on the weekend
Week 5	Intonation: listing, yes-no questions, OR-questions, WH-questions	Review and practice on intonation Progress test 3 Texts for more practice: 2 short stories from American Anecdotes – Elementary

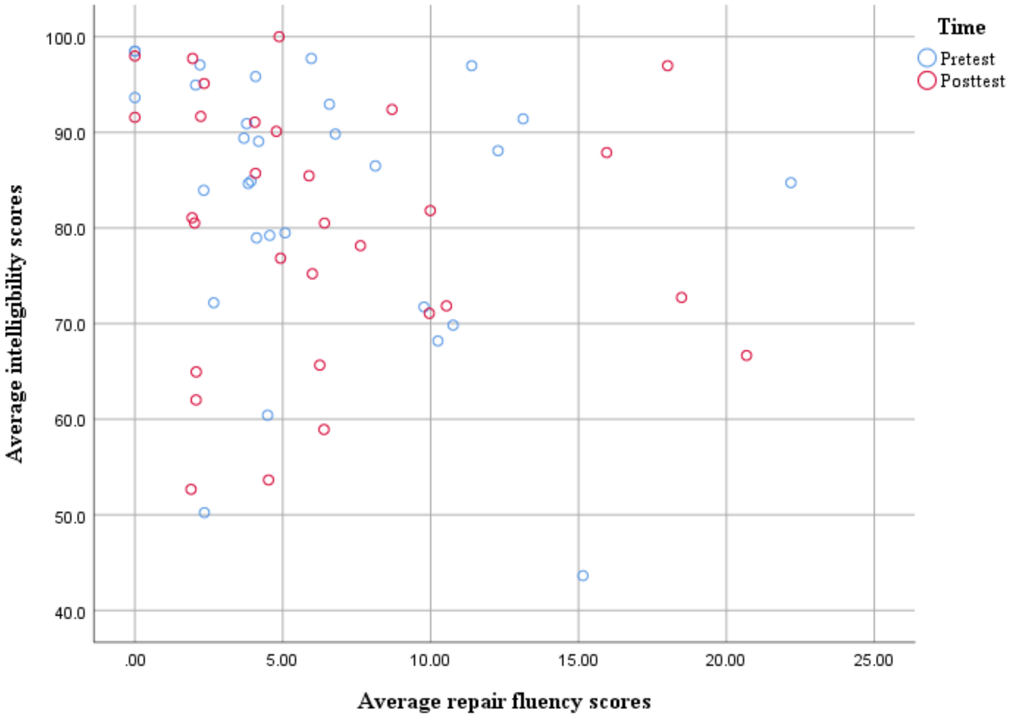
Week 6	Thought groups: rules for speaking in thought groups	Review and practice on thought groups Progress test 4 Speaking topics for more practice: (1) Would you prefer to live in the countryside or in a big city. Explain why; (2) Where would you like to travel if you have a chance
Week 7	Review on stress, de-stress, rhythm, intonation, and thought groups	Review and more practice with stress, de-stress, rhythm, intonation, and thought groups
Week 8	Wrapping up	

Appendix B. Scatterplots of Intelligibility and Fluency Scores.






Time
○ Pretest
○ Posttest
Average intelligibility scores
Average breakdown fluency scores
100.0
90.0
80.0
70.0
60.0
50.0
40.0
10.0
20.0
30.0
40.0
50.0
60.0
70.0


Time
○ Pretest
○ Posttest
Average intelligibility scores
Average repair fluency scores
100.0
90.0
80.0
70.0
60.0
50.0
40.0
.00
5.00
10.00
15.00
20.00
25.00

Appendix C. Utterance Fluency Predictors of Perceived Fluency Ratings.

Model	Predictor	Adjusted R^2	Change
1	Speech rate (SFlu)	.653	
2	Filled and unfilled pauses (BFlu)	-.033	
3	False starts, replacements, repetitions, and reformulations (RFlu)	-.009	
4	Speech rate (SFlu) and filled and unfilled pauses (BFlu)	.643	*F*(1, 56) = .17, *p* = .84
5	Speech rate (SFlu) and false starts, replacements, repetitions, and reformulations (RFlu)	.653	*F*(4, 54) = 1.02, *p* = .41
6	Speech rate (SFlu), filled and unfilled pauses (BFlu), and false starts, replacements, repetitions, and reformulations (RFlu)	.640	*F*(4, 52) = .90, *p* = .47