SILC Journal

Volume III

2024

Trials and Tribulations of Recruiting and Training Raters

Bertram Allan Mullin
Graduate College of Education, Temple University Osaka
bertram.mullin@temple.edu

Being contextualized around the process of training and recruiting raters, this paper provides recommendations to future researchers tasked with selecting and training raters. The paper reflects on research conducted with 21 Japanese adult participants who received computer-assisted pronunciation training treatments for four months over Zoom tutoring sessions. Eight raters and I reviewed three months of recorded data to track whether participants reached their individual pronunciation goals following treatments.

This paper covers the recruitment of raters for the project as well as my reflections on working with raters. Due to the need for more rater guidelines and strategies in rater use within recent studies, such data will prove useful for those who are skeptical about using human raters in their future research projects. Thus, the purpose of this paper was to inform other researchers of challenges that can arise when recruiting with a focused topic on recruiting, training, and working with human raters; and methods of the application of strategies for training raters effectively based on the literature.

本論文は、評価者のトレーニングと採用の過程を中心に説明しており、評価者の選出及び訓練を任されている将来の研究者に向けての提案をするものである。本論文では、4 か月間に渡り 21 名の日本人成人参加者を対象に実施した Zoom 上でのコンピュータを用いた発音トレーニングの個別指導を振り返る。８名の評価者と著者は、トレーニング実施後に参加者が各自の発音目標を達成したかを追跡するため、３か月間の録音記録を再考した。本論文ではこのプロジェクトの評価者採用及び、評価者との協力についての著者の省察も含める。最近の研究における評価者の採用には、より多くの評価者用ガイドラインと戦略の必要性が求められているため、このようなデータは今後の研究プロジェクトにおける人的評価者の採用に関し懐疑的である研究者にとっては有益であると考える。従って、本論文の目的は、採用、トレーニング、および人的評価者との協力に焦点を当てたトピックを採用する際に発生する可能性のある課題の報告、及び 文献に基づいて評価者を効果的に訓練するための戦略を適用する方法に関する情報を提供することである。

Introduction

In 2021, I recruited eight raters for a project that utilized computer-assisted pronunciation training analytical tools to coach 21 adult Japanese participants who wanted to improve their pronunciation using American English for professional and recreational reasons. The manuscript is under preparation for journal submission in 2024 (Mullin, 2024). While there are different forms of pronunciation (e.g., Australian, British, Canadian, etc.) and all are relevant, the phonemes taught in Japan are from American English. Researchers might want a short guide that helps recruit and train raters for their project. It is important to recognize that some literature on rating pronunciation exists in the field of second-language acquisition (SLA), and a few of the texts will be discussed in this paper (e.g., Kang et al., 2018; Winke & Brunfaut, 2021; Yan & Chuang, 2022; Youn, 2014). Moreover, Koizumi et al. (2017) highlighted the importance of precise training needed for raters. Lastly, it is beneficial to discuss Bond et al. (2020), who investigated rater bias and outlined detailed statistical analysis on calculating rater bias based on consistency measurement tools.

Arguably, the aforementioned literature does not concentrate strictly on strategies to recruit and train raters, but rather focuses on broader areas of SLA. For example, Bond et al. (2020) delves into statistical analysis to measure consistency and does not discuss rater recruitment. While Isaacs and Trofimovich (2017) wrote about rater recruitment, none of the researchers mentioned recruitment methods or how to maintain raters throughout a study.

Nonetheless, the lack of specific literature on recruiting and training raters who score pronunciation of learners has led to a gap in training strategy texts.

Thus, this essay sets out to accomplish the following goals: (a) inform other researchers of challenges that can arise when recruiting with a focused topic on recruiting, training, and working with human raters; and (b) apply strategies for training raters effectively based on the literature.

Literature Review

Among the literature on training raters, contrasting views were evident. Youn (2014) and Bond et al. (2020) both pointed out challenges raters might pose, but that even strict raters were at least consistently strict, thereby not biased when rating. Meanwhile, Koizumi et al. (2017) pointed out how strict and lenient raters might be biased in other ways. This literature review will examine each point.

First, Youn (2014) discussed how to train raters, and examined rater reliability to analyze pronunciation under the guise of speaking tasks. The project had 12 raters (three men and nine women) with master's degrees rate audio data from 102 university English language learners (ELLs) who were roleplaying. While five raters spoke English as a foreign language, seven were from English-speaking countries. The point of Youn's study was to determine whether consistency was a factor when raters' skills differed. The raters had the same training materials, including transcripts of the recorded data. After rating 102 participants, the data revealed that all raters scored consistently, meaning that if one rater usually marked an average score of five, then that rating remained consistent through scoring, according to Youn (2014).

Other researchers who used raters in past literature have also referred to Rasch statistical analysis to analyze rater reliability, including Koizumi et al. (2017). The researchers measured rater-reliability to determine whether raters were different in measurement. Specifically, in order to test rater-reliability Koizumi et al. (2017) measured three facets of reliability. First, if the scores were similar. Second, if the scores were consistent. Third, if the scores were biased. The study had 13 trained raters score 648 participants (high school students) who took the Global Test of English Communication Computer-Based Testing exam. The raters were required to have experience testing ELL-speaking ability. Raters were trained and observed for one week following a rater certification test. Then, raters embarked on trial ratings and received feedback after a second session. The researchers wanted to determine whether raters gave inconsistent scores for the same prompts during a second trial. The results indicated that raters scored differently than one another, as individuals bring their own characteristics, judgements, preconceived notions, and behaviors with them when scoring, according to the researchers. Thus, individual rater scores will most likely differ between raters. Koizumi et al. cautioned that researchers understand that some raters will rate harshly, and others leniently based on their own ideologies and beliefs.

Regarding training raters to score pronunciation specifically, Kang et al. (2018) wrote a handbook with several chapters dedicated to teaching pronunciation including aspects of phonology, defining phonetics, developing pronunciation, and training raters to score pronunciation along with several other topics in the handbook. Kang et al. stated that training reduced rater bias when rating pronunciation and added that rubrics for grading were pivotal in controlling prejudice. According to Kang et al., human raters can stereotype someone's intelligence based on accent. Kang et al. stated that untrained raters tend to judge a speaker on their accent rather than specific problems, such as a Japanese ELL stressing sounds that should not be stressed in an English word. Thus, the researchers cautioned that rubrics should not be too vague or misleading. Instead, the rubric should state exactly what is being rated, such as prosody (rhythm), and it should define terms, so raters know what to analyze rather than being able to rate based on their own biases.

According to Bond et al. (2020), bias exists in all human characteristics, which is noteworthy when using human raters for assessments. Moreover, Kang et al. stated that statistics should be used to support any statement that human-rater bias was avoided, such as

accent bias, and auditory or listener bias (i.e., when someone has a preconceived notion that a mistake from a person speaking a second language will be uttered at any moment).

Moreover, Winke and Brunfaut (2021) suggested selecting raters based on their familiarity with the subject matter being scored, thus experience with a subject was a strong qualifier. Another important aspect of the training process was to counter any rater bias toward language proficiency. Winke and Brunfaut stressed that scorers should be "clear, detailed, fair, and free from bias" (p. 301). In other words, a rater must completely understand the scoring system to avoid bias and maintain rater-reliability. The researchers (Winke & Brunfaut, 2021) suggested that minor pronunciation problems (e.g., an ELL produces *and* with an *o* at the end of the word) did not determine language proficiency as a qualifier for good pronunciation.

On qualifiers, Yan and Chuang (2022) recruited certified raters for an English placement test at a university in the United States. Raters received one semester of training to evaluate the evolution of raters on a longitudinal basis. In that period, the researchers found that even raters with regular training can score inconsistently. The researchers stated that there was a gap in literature that discussed rater training over an extended period of time. Furthermore, Yan and Chuang posited that often rater scoring was inconsistent, so even training does not remove rater bias entirely. On that note, Bond et al. (2020) suggested that statistical evidence will help researchers catch potential bias. Thus, Yan and Chuang used many-facets Rasch statistics analysis and concluded that the raters' inconsistent scores were due to rater opinions developing over time.

The past research discussed had a common trait of discussing consistency, fairness, and rater bias around training structure. Therefore, this project aimed to both advise and apply strategies based on the aforementioned studies above and my experiences on recruiting and training raters.

Background and Concerns Prior to Rater Selection

The project on pronunciation rating was part of the coursework leading to a dissertation in a doctoral program in applied linguistics. In the coursework, I learned how to recruit and train raters for a research project. For example, foundations in research and assessment courses that took place for approximately one year emphasized rater recruitment and training tactics and strategies. The instructors organized lessons around how to train raters and use Rasch analysis to confirm rater-reliability and to spot rater bias. A significant takeaway for the project I worked on later was to be detailed when outlining raters' expected tasks and when explaining the procedures for rating participants' pronunciation samples. Thus, the raters had rubrics, guidelines, example scores of data and comments, and tutorials available during training.

A faculty adviser mentored me throughout this project. The adviser's guidance included assistance in creating the pronunciation grading rubric, removal of unnecessary questions in the grading system prior to rater selection, and support in wording the 5-point Likert scale to ensure it was user-friendly for the rater. Throughout the project, I learned how to prepare for training raters and what to avoid. This advice was beneficial in maintaining most of the raters for 4 months, with only two leaving the project for distinct reasons addressed later in the essay.

At the outset of the project, participants and raters used aliases, which they chose based on a list of fruits. Moreover, the eight raters did not see the participants' names. Raters received recorded data files with three sets of voice excerpts from post-assessments, which were labeled with the letters A to J. Raters were required to listen to samples of sentences spoken by up to three of the 21 participants during a pronunciation pre-assessment. During the project, the raters were asked to compare these recordings to three monthly post-assessment recordings. Details regarding how raters were trained to complete these tasks are provided below as a means of demonstrating that the primary concern during the training process was avoiding the possibility of rater bias.

To evaluate rater bias in this study, Bond et al.'s (2020) and Kang et al.'s (2018) texts were adopted to develop assessment content, which highlighted potential rater bias.

Specifically, the texts helped in the production of a 5-point Likert scale to score 13 prompts on the vocal production of pronunciation. Some questions were purposely worded negatively (e.g., "The participant said the sentences flat and almost nothing was stressed.") because Bond et al. (2020) pointed out that such statements could ensure that the raters were careful with their scoring. The rating process required raters to listen to audio recordings from participants saying sentences and score the participants from 1 (*strongly disagree*) to 5 (*strongly agree*) based on suprasegmental (intonation, structure, and stress) and segmental (consonants and vowel sounds, pauses) features of pronunciation. Raters had a reference to compare the recorded data: English speakers from America (two male and two female) who were recruited for the project as voice actors. Each rater received two recordings of the two American male speakers saying the same sentence as a male participant. If the participant was female, the rater received two recordings of American women saying the same sentence as a female participant.

To determine whether the raters judged participants' pronunciation proficiency on minor issues, and reveal potential bias, the raters briefly commented on the data recordings. Comments helped to better understand their rationale and scores, as shown in the example data in Appendix A. The raters evaluated the recordings from participants based on the scoring rubric by entering their scores into an Excel sheet. Aside from comments removed from the study that were offensive and unhelpful, a sample of the comments provided by raters are available in Appendix A.

In addition to the comments, two other methods assisted in determining whether rater bias was a factor. The first method was through statistical evidence, as Bond et al. (2020) suggested. That is, many-facets Rasch analysis was used to generate a Wright map, which indicated where the participants scored from the lowest to the highest scorer based on rater scores. The second method was a participant measurement report where Rasch analysis helped determine whether the participants were scored consistently by raters, i.e., I could detect if a rater consistently scored participants harshly or leniently based on the statistics generated with Rasch analysis, which were used in the results of the project that this paper is based on.

After running Rasch analysis, the data were cross-referenced with a rater measurement report, which revealed that Raters 3 and 8 were slightly misfit. In this context, those two rater scores were not always consistent, which was a concern. Interestingly, raters 3 and 8 were less experienced as ELL instructors than other raters. Thus, the statistical analysis was helpful because it showed that there were few inconsistencies in rater scores, which means that most of the raters were statistically consistent in scoring; thus, the variance between rater scores was statistically insignificant. Most importantly, overall consistency revealed a sign of lower rater bias. I could determine such information through Rasch analysis, as it measures categorical data, such as Likert-scale ratings, to analyze participants' skills by assessing item difficulty (Bond et al., 2020).

As mentioned at the beginning of this section, rater comments were gathered as another method to detect rater bias, as recommended in Kang et al. (2018). For example, one rater made unprofessional comments about a participant based on their accent, for example, calling the participant "hopeless." The comments helped me find areas where raters were judging minor problems and considered them proficiency factors, such as the one mentioned where the rater in question scored a participant as low as possible based on the participant's accent.

Consequently, the statistical evidence did not show inconsistency in rating, so the rater comments assisted in the realization that one rater was indeed being biased, as the individual was personally attacking a participant based on their accent and minor mistakes, such as adding an *o* at the end of the word *and*. Such a rationale was beyond the rubric that the raters were using. If the rater was scoring for something beyond the rubric, the comments were available to assess that the rater was potentially biased and basing a score on a minor issue, such as adding *o* to the end of the word *and*, rather than basing a score on what was asked of the rater in the scoring process. While one could argue that the mistake was a pronunciation issue, scoring someone with zero scores on every measure due to a single mistake was too

harsh. Thus, the rater's level of harshness became a concern, which I will discuss more in the challenges section.

Recruitment Process and Training

In October 2021, I posted in a social media group dedicated to teachers of ELLs in Japan explaining that raters were needed for a pronunciation project that lasted approximately 4 months. The post specified that raters should have taught English as a foreign language for at least 1 year, with some experience teaching pronunciation, or they should have studied to teach English and taught it for at least 1 year. The rationale for selecting second language instructors was due to the participants being ELLs in Japan; instructors would be familiar with participants' intelligibility. Rater selection was based on their skills and abilities. Someone experienced in teaching ELLs was well-suited to be a rater and thus qualified. Youn (2014), Koizumi et al. (2017), and Yan and Chuang (2022) recruited raters based on the individual's experiences and qualifications as well.

Several people initially responded to my post, all of whom were qualified, having at least 1 year of experience teaching ELLs. These potential raters were first informed that their tasks would involve scoring anonymous adult Japanese learners' sentences after each post-assessment, for a total of three times unless withdrawn from the study. The raters were also informed they would be paid 3,000 yen for their participation. Each rater scored three participants during each assessment on a 5-point Likert scale for each of the 13 questions related to how well the participants pronounced diphthongs, vowel utterances, and consonants (segmental features) as well as their stress, rhythm, and intonation (suprasegmental features). Raters first participated in Zoom training sessions, after which they received video tutorials and written and audio instructions on how to use the Excel file and score participants. Example scores and comments can be found in Appendix B. Raters were provided training materials in multiple mediums from videos and emails to audio versions of the emails, so that raters could access instructions most appropriate to their learning style; my mentors had stressed the importance of the researcher's role in ensuring raters clearly understand task expectations. The scoring process for three participants took approximately 1 hour to score three participants per rater. Moreover, each rater signed consent forms that stated that they had the right to withdraw at any time and that their data and scores could be deleted upon request.

Challenges

After recruitment, I met with the raters individually on Zoom as needed. During the first Zoom session, raters reviewed the Excel file with a grading rubric, which contained the Likert scale and scoring rubric. Some of the raters did not like using Excel. One rater, for example, printed the file out and handwrote their scores. While scoring the first post-assessment, another rater had difficulty using Excel and decided to drop out of the study after rating their first three participants. I had to recruit another rater who had previously shown interest and expressed no issues using Excel. Although the rater who dropped out was the most experienced instructor of all raters, with over 30 years of experience teaching ELLs, the rater said that they were not comfortable scoring participants. As the rater explained, listening to the data and using the Excel file to score the participants was too complicated. Unfortunately, the rater was unwilling to discuss what could be done to ease their frustration and continue in the study; they simply said they felt unqualified and no longer wanted to participate. However, the rater was fine with their scores being used as data in the research.

Because the raters lacked experience rating participants, the raters expressed feelings of inadequacy and self-doubt during our Zoom sessions and email correspondence. To ease their doubts, I provided consistent communication and advice throughout the process. Two raters who expressed insecurities were Japanese instructors of English. Both felt less qualified than raters from America who rated Japanese learners. Acknowledging that other raters doubted themselves seemed to alleviate these concerns.

During the second post-assessment, a rater commented that one participant was "hopeless" and gave the lowest possible scores. Their other comments toward that participant

were even more severe and distasteful. Another rater scored the participant to ensure no bias, even on my part, occurred, then I compared the new scores to my own. Only the rater in question scored the participant with the lowest possible scores. However, the scores from the other two raters were on the low end and despite the comments, the rater's statistics were consistently harsh for the other participants' scores, thus did not reveal bias. Nevertheless, my university advisors said removing the comments and replacing the rater was understandable. My decision to remove the rater was due to the risk of a repeated similar incident. Although this rater was knowledgeable about good pronunciation, their comments demonstrated accent and listener bias, which aligned with Kang et al. (2018), who said accent biases exist and was a factor that Winke and Brunfaut (2021) cautioned researchers of when scoring. Removing the rater was surprisingly not an issue. I told the rater that enough data and ratings were collected and sent their payment. The rater said the process was fun and wanted to rate participants in the future. Rather than recruiting another rater, the remaining raters volunteered to score an extra participant when asked.

The remaining raters described the rating process as enjoyable. All except one showed interest in future rating-related projects. At first, no raters were confident about rating, but repeated assurance that the statistics showed consistent ratings helped the raters see that there were no issues with their scores. Only one rater scored a bit randomly. That rater was also the least confident and was the one who did not want to participate again.

Reflection

Aside from keeping up with rater scores, the most difficult part about training raters for me was the need to accept that some raters might withdraw. Future researchers should plan for this issue by having alternate raters available. Replacing the rater who quit took time away from other project responsibilities. It is advisable to accept that no matter how much training and support is offered to raters, certain individuals will quit because of internal doubts or external factors such as personal life situations.

Another point worthy of reflection is that a future study could benefit from considering the experience level of the rater, which will ensure that an individual has the required skills to be an effective rater. This study used social media platforms to scout for raters, but there is also the option of asking colleagues to become raters. However, setting higher standards makes finding raters more difficult, and the task can become time-consuming. Nevertheless, as my mentors explained, considering the quality of raters is important to ensure the integrity of data, and that one should ensure that participants receive fair and consistent assessments. Notably, the two raters with the least amount of experience teaching English presented the most problems: (1) the rater with the least experience, exactly 1 year teaching ELLs, was the most inconsistent rater according to Rasch statistics; and (2) the rater with the second-shortest time teaching was the one removed from the study due to inappropriate comments about a participant's accent. In hindsight, more experienced teachers were more consistent scorers, so whether 1 year of experience is enough to qualify someone as a scorer is open to debate. Thus, it will be useful for future researchers to search for experienced raters.

Conclusion

The purpose of this paper was to accomplish two goals. The first goal was to inform other researchers of challenges that can arise when recruiting, training, and working with human raters. Among the challenges, I discussed how raters could suddenly drop out of a study due to self-doubt. Other raters might have to be removed from a study due to their unprofessional behavior. One rater used handwritten notes rather than the provided Excel file. Having a preformatted Word document could prevent such issues and make the rating more accessible to those with various technology skill levels. The major challenge of training was having to replace raters. However, future researchers might consider the importance of having alternate raters on standby. The second goal was to apply strategies for training raters effectively, using the guidance of past research and seeking advice from my mentors was advantageous in the rating process. Specifically, the mentors' input, such as using a grading rubric, helped as I built strategies to effectively train raters. My advice is that researchers remember that the

rating process is not a single-person job. Raters require mentorship, as do the researchers themselves. I am grateful for the existing guidance and highly recommend the texts mentioned for further reading. Bond et al. (2020) is especially useful for rating strategies in general, as well as for understanding how to conduct Rasch analyses and interpret the statistics. For tips on rating pronunciation and limiting rater bias, Kang et al. (2018) is a great resource. For a future study, Yan and Chuang (2022) provided useful insights on rater-reliability and training raters. Moreover, Koizumi et al.'s (2021) training process is worth application, especially having raters do trial ratings and providing feedback. That process was unique in maintaining rater-reliability through their study and is noteworthy for future research. Additionally, various training materials, as Youn's (2014) provided, will ensure that a rater has the tools to score participants. Finally, remember that raters are also human beings with their own problems and challenges, so be compassionate.

Acknowledgements

I am grateful to the dedicated raters involved with this project. It humbles me to have received help from such great professors as Dr. David Beglar, Dr. Nathaniel Carney, and the Temple University instructors. Moreover, thank you to Kathleen Wilson Legg for editing the first draft and Dr. Peter Ferguson for editing and providing time and feedback on additional drafts. Lastly, thank you to the hard-working editors and reviewers at the SILC Journal. Any mistakes found in the final version are that of the author's sole responsibility.

References

Bond, T., Yan., Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Isaacs, T., & Trofimovich, P. (Eds.). (2017). Second language pronunciation assessment: Interdisciplinary perspectives (vol. 107). *Multilingual Matters / Channel View Publications*. http://www.jstor.org/stable/10.21832/j.ctt1xp3wcc

Kang, O., Thomson, R., & Murphy, J. (2018). *The Routledge handbook of contemporary English pronunciation*. Routledge. https://doi.org/10.4324/9781315145006

Koizumi, R., Okabe, Y., & Kashimada, Y. (2017). A multifaceted Rasch analysis of rater reliability of the speaking section of the GTEC CBT. *J-Stage*, *28*, 241–256. https://doi.org/10.20581/arele.28.0_241

Mullin, B. A. (2024). English as a foreign language education: HVPT treatments [Unpublished manuscript]. Graduate College of Education, Temple University. https://www.tuj.ac.jp/grad-ed

Winke, P., & Brunfaut, T. (Eds.). (2021). *The Routledge handbook of second language acquisition and language testing* (1st ed.). Routledge. https://doi.org/10.4324/9781351034784

Yan, X., & Chuang, P-L. (2022). How do raters learn to rate? Many-facet Rasch modeling of rater performance over the course of a rater certification program. *Language Testing*, *40*(1), 1–27. https://doi.org/10.1177/02655322221074913

Youn, S. J. (2014). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(2), 199–225. https://doi.org/10.1177/0265532214557113

Appendix A: Examples of Rater Comments

Below are samples of rater comments pulled randomly from three of 21 participants, one per group.

Group	Participant	Assessment	Rater	Rater comment
A	1	1	7	The participant seemed to have acceptable intonation and pronunciation and adequate control.
A	1	2	4	Need improvement: “th” and ending /r/ sounds. Improved compared to the pre-assessment. All areas are showing improvement from the sample pre-assessment.
A	1	3	2	I can hear his improvement from the pre-assessment. However, he still is having some troubles making natural pauses, and with the stressed and unstressed parts.
B	8	1	8	The participant only has some control over vowels and makes /v/ or /f/ sounds when pronouncing the diphthong /əʊ/ sentence: “Though the coach has a goal, he isn’t very focused.”
B	8	2	5	No improvement between the pre-assessment and assessment.
B	8	3	3	Compared to the pre-assessment recording, the intonation and stress of /d/ sounds are getting better. Some of the participant’s voice pitch has a unique katakana sound, but overall improvements.
C	20	1	6	Her weakest point seems to be pauses.
C	20	2	3	She stressed almost every word, but it's still understandable what she says. Comparing to ES samples, her sentences sound more flat.
C	20	3	7	The participant’s pronunciation of the double /l/ in traveling was a little weak but normal. The /ow/ in now was a little weaker than the other /ow/ words. As for the vowel and “said,” it sounded more like “si.” I could not hear a clear /d/ which sounded more like a /t/. However, the participant should be commended for pronouncing the final “er” in December very clearly.

Appendix B: Grading Rubric

Below is a copy of the grading rubric with an example rating and comments provided to the raters.

Grading Rubric Prompts	Example Score
Compared to the 1st English speaker, the participant's segmental (consonants, vowels, vowel diphthong) stress had ***good*** pronunciation control. *1 = strongly disagree, 2 = disagree, 3 = somewhat agree, 4 = agree, 5 = strongly agree*.	Leave blank
Consonants compared to English Speaker 1	3
Vowels compared to English Speaker 1	2
Vowel Diphthong compared to English Speaker 1	5
Compared to the 2nd English speaker the participant's segmental (consonants, vowels, vowel diphthong) stress had ***good*** pronunciation control. *1 = strongly disagree, 2 = disagree, 3 = somewhat agree, 4 = agree, 5 = strongly agree*.	Leave blank
Consonants compared to English Speaker 2	2
Vowels compared to English Speaker 2	1
Vowel Diphthong compared to English Speaker 2	2
Compared to the PreTest sample, pronunciation improved overall	1
Suprasegmental (sentence structure, stress, and intonation) stress had ***good*** pronunciation control	3
Unstressed sounds or weak vowels (e.g., "i" in "Africa") were unstressed (see /ə/ in the IPA Pronunciation English app used for the test for a full list of syllables).	3
The participant said the sentences flat and almost nothing was stressed.	2
The participant stressed almost every syllable in the sentence	2
Silent pauses in the sentences were natural in my opinion (and based on the 1st English speaker sample).	3
Silent pauses in the sentences were natural in my opinion (and based on the 2nd English speaker sample).	3
Total	32
Example comments	The participant only had some control or little to no pronunciation control. Their voice volume was low, strengths were confidence and self-corrections, and problem priorities were L and V phonemes.

崇城大学 SILC 紀要

SILC Journal

2024 年 2 月 21 日

編集・発行　崇城大学 Sojo International Learning Center
〒860-0082 熊本市西区池田 4 丁目 22-1

印刷・製本　崇城大学 出版センター