The Relationship between Japanese EFL Learners' Perceived Fluency and Temporal Speech Measures in a Read-Aloud Task

Aki Tsunemoto
Concordia University

Pakize Uludag
Concordia University

Kim McDonough
Concordia University

Talia Isaacs
University College London

This study examined the relationship between holistic rater judgments of second language (L2) speech fluency (i.e., perceived fluency) and temporal measures of fluency (i.e., utterance fluency) in a read-aloud task. Sixty-three L2 English Japanese secondary school students were audio-recorded while carrying out a 69-word read-aloud task. Eleven L2 English-speaking instructors rated the speech for perceived fluency, and the speech samples were analyzed for utterance fluency. The linear regression model revealed that articulation rate and clause-internal pauses significantly predicted perceived fluency. Findings are discussed in relation to the use of read-aloud tasks for the teaching and assessment of L2 speech fluency.

本研究では、音読タスクにおける第二言語音声の流暢性に関する総合的評価と、スピーチの言語的特徴の関係性を調査した。高校生の日本人英語学習者63名が、69語の音読タスクを実施する様子を録音した。11人の言語指導経験のある第二言語英語話者が、スピーチの流暢性について評価した。更に、スピーチを言語的特徴について分析した。重回帰モデル解析の結果、発声速度と節内のポーズが流暢性の重要な予測変数として算出された。これらの成果に基づき、音読タスクを利用した第二言語における流暢性の指導や評価について教育的な示唆を行う。

Keywords: English as a Foreign Language; fluency; second language speaking; speech perception; read-aloud task

Whereas fluency in a broad sense is often equated with general oral proficiency, fluency in a narrow sense refers to the temporal fluidity of speech (Lennon, 1990), specifically whether it is smooth and rapid (De Jong, 2018). One goal of fluency research has been to understand the relationship between utterance fluency (i.e., speech features), and perceived fluency, which captures raters' impressions of utterance fluency (Segalowitz, 2010). To gain insight into this relationship, utterance fluency has been measured in terms of speed fluency (e.g., speech rate), breakdown fluency (e.g., duration and frequencies of pauses) and repair fluency (e.g., frequency of self-corrections and repetitions) (Tavakoli & Skehan, 2005) while perceived fluency has been assessed through holistic rater judgments. Prior studies of fluency during spontaneous speech reported a positive relationship between perceived fluency and speed fluency measured as speech rate (e.g., Magne et al., 2019) and mean length of run (MLR: e.g., Derwing et al., 2004; Kormos & Dénes, 2004; Trofimovich, et al., 2017). In contrast, perceived fluency has been negatively associated with breakdown fluency measured as the frequency and durations of silent pauses (e.g., Rossiter, 2009), pauses within clauses (e.g., De Jong & Bosker, 2013; Kahng, 2018; Suzuki & Kormos, 2020), and pauses between clauses (Saito et al., 2018). Finally, perceived fluency has shown both positive (Magne et al., 2019; Saito et al., 2018) and negative (Kormos & Dénes, 2004) relationships with repair fluency.

Although the relationship between perceived and utterance fluency has been widely examined in spontaneous speech, less is known about their relationship during read-aloud tasks, which are commonly used for both English proficiency testing and pedagogical activities. Several high-stakes English proficiency tests use read-aloud tasks, often combined with automated scoring, as part of their speaking assessment (e.g., Duolingo, EIKEN, GTEC, Pearson Test of English Academic [PTEA]), including new tests developed in response to the COVID-19 pandemic (e.g., TOEFL Essentials, Isbell & Kremmel, 2020). In Japan, English learners may take these tests for admission to foreign universities or for immigration purposes. Among such tests, EIKEN, which includes a read-aloud task for most grade levels, is taken by three million people each year as a gatekeeping measure to demonstrate English proficiency for post-secondary education and employment in Japan (EIKEN, n.d.). Furthermore, in second language (L2) classrooms, read-aloud tasks have been included in diagnostic pronunciation assessment to identify learner needs and create individualized instruction (Celce-Murcia et al., 2010). In Japan specifically, where people predominantly use only Japanese on a daily basis (Coulmas & Watanabe, 2011), instructors often implement controlled tasks such as reading aloud from textbooks (Uchida & Sugimoto, 2018). Most teachers in Japan tend not to incorporate extemporaneous speech tasks into their classes and often use scripted tasks when evaluating speaking performance (for review, see Koizumi, 2022).

Unlike spontaneous speech, read-aloud tasks do not require speakers to conceptualize message content. Instead, they need to parse the textual information, encode phonological information, and execute the planned phonetic information into sounds using physiological mechanisms. Although read-aloud tasks require this complex processing, they do not require speakers to pre-plan content, retrieve words, or build grammatical structures as in spontaneous speech tasks. As a result, a speaker may produce more regulated speech patterns (Laan, 1997) and speak faster with fewer hesitations (Trofimovich, et al., 2017) during read-aloud tasks than spontaneous speech. The lower variability in speaker performance is conducive for machine scoring, making the read-aloud task attractive as a time-efficient, reliable, and inexpensive test item that can be scored automatically (Isaacs, 2018). Nevertheless, in languages like English with poor sound-symbol correspondence, read-aloud tasks may still pose challenges for speakers, such as mispronouncing words that have irregular written forms or hesitating before unfamiliar words (Hayes-Harb et al., 2010), and these challenges may influence rater perceptions of their fluency.

In light of the role of read-aloud tasks in L2 assessment and classroom practices in English L2 settings, it is important to investigate speech characteristics that are perceptually salient to L2 English speakers. The few prior studies that included read-aloud tasks with L2 Dutch and L2 French speakers found that perceived fluency was positively associated with speed and repair fluency but negatively related to breakdown fluency (Cucchiarini et al., 2002; Trofimovich, et al., 2017). However, both studies elicited evaluations of perceived fluency from first language (L1) speakers of the target language. Prior studies of perceived fluency during spontaneous speech found that both L1 and L2 English raters were influenced by speed and clause-internal pauses, but only L1 raters were sensitive to clause-external pausing (Magne et al., 2019; Saito et al., 2018). Little is known, however, about whether these utterance fluency measures are equally important for L2 English speakers when assessing L2 fluency through a read-aloud task. Due to globalization, most English speakers are now L2 speakers (Pennycook, 2020) and many work as instructors and language test examiners (Carey et al., 2011), which highlights the need for further research to elicit their perceptions of fluency. Against these backdrops, the current study examines the relationship between L2 English-speaking instructors' perceptions of fluency and temporal measures of Japanese English as a foreign language (EFL) students' read-aloud task performance. The research question was as follows:

> RQ. What temporal measures of speech fluency (i.e., utterance fluency) are associated with L2 English-speaking teachers' holistic fluency ratings (i.e., perceived fluency) during a read-aloud task?

Method

L2 Speakers

As part of a larger study, L2 speech samples were elicited from 63 secondary school students in Japan (45 males, 18 females, M_{age} = 16.4, *SD* = 0.6). All students and parents were L1 Japanese speakers except for one Japanese-Korean bilingual student. The students began studying English around the age of 10.5 years (*SD* = 3.1) and except for the bilingual student, they had no experience living in English-speaking countries longer than a month. All but eight students self-reported their most recent EIKEN Grades (*range* = Grade 1–4), 80% of whom reported achieving Grade 2nd, Pre-2nd, or 3rd. Their English classes primarily targeted reading and writing skills, and speaking activities usually involved reading words and sentence aloud from a textbook, occasional paired or group discussions, and bi-weekly sessions with an assistant language teacher. Some students voluntarily participated in after-school English conversation groups.

Task and Speech Recording

During an individual session with the first researcher (15 minutes), the students completed a read-aloud task based on a passage from the Speech Accent Archive (Weinberger, 2015; see Appendix A). The 69-word passage was selected because it contained all possible English sounds for eliciting the students' phonological encoding skills (Cucchiarini et al., 2002). Each student was given the passage and were asked to read it silently within one minute. After having the opportunity to ask about the meaning or pronunciation of any unfamiliar words, each student read the passage aloud while being audio-recorded. The audio-recordings, which ranged in length from 22 to 47 seconds, were trimmed by removing initial pauses and hesitations and normalized for peak intensity. The recordings were organized into three lists with different orders to limit the possibility of ordering effects.

Raters and Rating Procedure

Reflecting our focus on L2 English-speaking raters, we purposefully recruited L2 English speakers who had teaching experience. To ensure consistency in their familiarity with the Japanese language (Carey et al., 2011), we recruited raters who had never lived in Japan and did not speak Japanese. Through convenience sampling, 11 L2 English raters (10 females, 1 male) with experience teaching English to L2 learners (*M* = 5.8 years, *SD* = 4.0) were recruited. They were adults (M_{age} = 31.4 years, *SD* = 6.5) enrolled in or recent graduates of Education programs at an English-medium Canadian university. As degree seeking students, they had met the university's minimum English language requirement for admission without additional language instruction, which was a TOEFL iBT score of 90 (or equivalent). On a background questionnaire (Appendix B), they reported varied L1 backgrounds, including Chinese, Dutch, Farsi, Polish, Portuguese, Russian, and Vietnamese. They all reported having normal hearing, and nine reported having previously taken a phonology course. They estimated the percentage of time that they used English in their daily life on a scale of 0 to 100% for both speaking (*M* = 69.1%, *SD* = 24.3) and listening (*M* = 74.6%, *SD* = 21.2). When asked to self-report familiarity with L2 accented English on a percentage scale[1] (Tsunemoto et al., 2021; 0 = *not at all*, 100 = *very familiar*), the raters indicated that they were very familiar with L2-accented English (*M* = 77.8%, *SD* = 17.2), but not very familiar with Japanese accents specifically (*M* = 27.3%, *SD* = 26.1). None of the raters had previously lived in

Japan and they reported spending little time in their daily lives' interacting with Japanese speakers (*M* = 9.1%, *SD* = 16.1) when the study was carried out.

The raters scheduled individual rating sessions (60 min) with the first or second researcher held in a quiet room on a university campus in Canada. All 11 raters evaluated the entire 63 speech samples on a computer connected to a headset using 9-point Likert-type fluency scales (1= *not fluent at all*, 9= *very fluent*) in accordance with L2 speech fluency research conventions (e.g., Suzuki & Kormos, 2020). In line with previous studies that have revealed highly consistent fluency ratings among raters (e.g., Trofimovich, et al., 2017), raters were asked to judge how smooth the oral delivery was while focusing on temporal features (speech rate, fillers, pauses) in the speech (e.g., Kahng, 2018). After completing three practice ratings, they had opportunities to ask about the speech samples or rating scale. They were instructed to listen to an entire speech sample before providing a fluency rating. Raters were randomly assigned to one of three presentation orders to avoid possible ordering effects. The internal consistency of the raters' perceived fluency ratings was assessed by Cronbach's alpha, which was .91. Interrater reliability was assessed through two-way random, agreement, average-measure intraclass correlation coefficients. The obtained value was .88, which revealed acceptable rater agreement (Field, 2018; Kahng, 2018). As the consistency exceeded the threshold values of .70–.80 (Larson-Hall, 2010), fluency ratings were averaged to derive single mean scores for each speech sample.

Speech Analysis

The speech samples were analyzed for six temporal measures of speech that reflect speed fluency, breakdown fluency, and repair fluency. Although prior research has used a number of utterance fluency measures (e.g., Tavakoli, et al., 2020), we selected measures from previous studies with EFL Japanese speakers (e.g., Saito, et al., 2018) or read-aloud tasks (e.g., Cucchiarini et al., 2002). For speed, articulation rate was calculated as total syllables divided by total phonation time (subtracting the total silent pause duration from the total speech duration, Prefontaine et al., 2016). Four pause measures were used to assess breakdown fluency (MLR, clause-external, clause-internal, and filled pauses). MLR (total syllables/utterances produced between silent pauses) has been examined as speed measure (Prefontaine, et al., 2016), but we considered the variable as breakdown measure as it incorporates pauses and may represent a speaker's hesitation (Towell et al., 1996). As for pauses, any silences longer than 200ms were operationalized as pauses. A shorter duration than De Jong and Bosker's (2013) recommended cut-off (250ms) was used because read-aloud tasks require shorter periods to produce speech as compared to spontaneous speech (e.g., Cucchiarini, et al., 2002). Silent pauses were manually coded using Praat (Boersma & Weenink, 2017) with the assistance of automated silence detection. Pauses were categorized as either clause-external or clause-internal to examine relative contribution of pause location to perceived fluency ratings (Bosker et al., 2013; Kahng, 2018; Saito et al., 2018). Filled pause frequency was obtained as total number of dysfluencies (e.g., uh and um) divided by total phonation time (Bosker et al., 2013). Repair fluency was operationalized in terms of the repair ratio, which is the total number of dysfluencies (e.g., self-corrections and repetitions) divided by the total number of syllables in a passage from the Speech Accent Archive (Weinberger, 2015) to obtain a standardized measure that are comparable across speakers. A subset of the data (25%) was coded by the first researcher and an independent rater. Two-way mixed, agreement, average-measure intraclass correlation coefficients revealed high agreement values for clause-external pause frequency (0.97), clause-internal pause frequency (0.92), filled pauses (1.00) and total dysfluencies (0.88). Having established coding reliability, the remaining speech samples were coded by the independent rater.

Results

The descriptive statistics for the perceived fluency ratings and utterance fluency measures are provided in Table 1. The raters provided a wide range of L2 fluency ratings (3.2–7.8 on a 9-point scale), with a mean score slightly above the scale midpoint (*M* = 5.3). Overall, L2 speakers produced all types of utterance fluency measures, but filled pauses and repairs occurred less frequently.

Table 1

Descriptive Statistics for Perceived Fluency and Utterance Fluency

Variables			*M*	*SD*	*Min*	*Max*
Perceived Fluency		Raters' ratings	5.30	1.08	3.18	7.82
Utterance Fluency	Speed	Articulation rate	3.21	0.38	2.17	4.52
	Breakdown	Mean length of run	5.75	2.07	3.00	13.80
		Clause-external pause frequency	0.29	0.08	0.08	0.47
		Clause-internal pause frequency	0.22	0.14	0.01	0.66
		Filled pause frequency	0.03	0.06	0.01	0.25
	Repair	Repair ratio	0.04	0.03	0.01	0.15

Half of the utterance fluency measures had skewness and kurtosis indices larger than ±2 and examination of the histograms suggested that the data were not normally distributed (Field, 2018).

Therefore, a nonparametric Spearman's rank-order correlations were obtained to determine the relationship between utterance fluency and perceived fluency (see Table 2).

Table 2

Correlations between Perceived Fluency Ratings and Utterance Fluency Measures

	1	2	3	4	5	6
Perceived Fluency	0.71**	0.67**	0.26*	−0.71**	−0.21	−0.23
1. Articulation rate	-	0.46**	0.28*	−0.41**	−0.16	−0.17
2. Mean length of run		-	−0.22	−0.87**	−0.26*	−0.25*
3. Clause-external pause frequency			-	−0.15	−0.11	−0.24
4. Clause-internal pause frequency				-	0.31*	0.37**
5. Filled pause frequency					-	0.54**
6. Repair ratio						-

Note. $^{*}p < .05$, $^{**}p < .01$.

Based on the correlation coefficients, MLR was dropped from further analysis because it was strongly correlated with clause-internal pauses[2]. The three remaining variables that reached the benchmark for a small correlation coefficient of ±.25 (Plonsky & Oswald, 2014) were selected for inclusion in a hierarchical multiple regression model: articulation rate, clause-external pauses, and clause-internal pauses. Regarding assumptions and model fit, tests of multicollinearity showed that the model all tolerance values were above .2, and no VIF values were above 10 (1.00 to 1.24). The Durbin-Watson statistic indicated good model fit (1.84). The normality of residuals was determined by (a) visual inspection of histogram, scatterplots, and P-P plots, (b) fewer than 5% of cases with standardized residuals greater ±2, and (c) Cook's distance and DfBeta values were less than 1 (Field, 2018).

Because previous research has identified the importance of speed fluency, articulation rate was entered first followed by the two breakdown fluency measures. As shown in Table 3, the first model with articulation rate was significant, but the second model with clause-external pauses and clause-internal pauses led to a significant F change and higher R^2 value.

Table 3

Summary of Hierarchical Regression Models for Raters' Ratings

Blocks	*R*	R^2	ΔR^2	ΔF	*p*
1. Articulation rate	0.66	0.43	0.42	46.66	.001
2. Clause-external pauses & clause-internal pauses	0.85	0.71	0.70	28.85	.001

Both articulation rate and clause-internal pauses were significant predictors of L2 raters' perceived fluency in the second model and they explained a combined 71% of the variance, R^2 = .71, *F*(3, 59) = 48.99, *p* < .001. (see Table 4).

Table 4

Summary of Predictor Variables for Regression Model with Blocks 1 and 2

Predictors	*B*	*SE B*	Β	95%CI		*t*	*p*
Articulation rate	1.87	0.27	0.66	1.32	2.42	6.83	.001
Clause-external pause	1.15	0.99	0.09	−0.82	3.12	1.17	.248
Clause-internal pause	−4.13	0.55	−0.55	−5.24	−3.02	−7.46	.001
Constant	1.71	0.72		0.26	3.15	2.36	.021

Discussion

This study examined which temporal measures of utterance fluency are associated with L2 English speakers' holistic ratings of students' perceived fluency during a read-aloud task. The positive relationship between articulation rate and perceived fluency is in line with previous read-aloud task studies that demonstrated a positive link between articulation rate (i.e., mean syllables per second excluding pauses) and L2 Dutch fluency ratings (Cucchiarini et al., 2002) or between MLR and L2 French fluency ratings (Trofimovich et al., 2017). Put simply, these EFL speakers were perceived to be more fluent if they produced more syllables per second when reading aloud. Additionally, perceived L2 fluency was negatively associated with clause-internal pauses. Although prior read-aloud research identified a negative association between perceived fluency and the duration and frequency of silent pauses (Cucchiarini et al., 2002), the current findings indicate that only clause-internal pauses predicted perceived fluency. When reading aloud, pausing at clause boundaries may have occurred when these EFL speakers were organizing words into meaningful chunks, which did not influence these raters' perceptions. However, when they paused within clauses, such as when hesitating to pronounce unfamiliar words, they were perceived to be less fluent.

An example of clause-internal pauses is provided in the excerpt below ([*] represents a 200ms or longer clause-internal pause). This student received a low fluency rating (3.18 on a 9-point scale) and her speech contained numerous clause-internal pauses. Even though the student had chances to check the pronunciation of the unfamiliar words before reading aloud, clause-internal pauses seem to occur before unfamiliar words (e.g., slabs, plastic, scoop). There were pauses before more familiar words (e.g., big, bags, train), which suggests that the student did not put words into chunks, such as noun phrases (e.g., a big toy frog, three red bags) or prepositional phrases (e.g., at the train station).

> S56: Please [*] call Stella. Ask her to bring [*] these [*] things with her from the [*] store. Six [*] spoons of fresh snow [*] peas, five thi-[*]-ck [*] slabs [*] of blue cheese, and [*] maybe a snack for her brother Bob. We also need [*] a small [*] plastic snake and [*] a [*] big [*] toy frog for [*] the kids. She can s-[*]-coop [*] these things into three red [*] bags, and we will go meet her [*] Wednesday at [*] the [*] train station.

Finally, in contrast to speed and breakdown fluency measures, repair fluency occurred relatively infrequently and did not predict perceived fluency, which is in line with previous studies that demonstrated small negative correlations between repair fluency and perceived fluency in L2 Dutch (r = –0.15, Cucchiarini et al., 2002) and L2 French (r = –0.24, Trofimovich et al., 2017).

The current study raises some potential implications for L2 instruction and assessment. Instructors may help students increase their articulation rate and decrease their clause-internal pauses by having them read the same text aloud repeatedly (Yoshimura & MacWhinney, 2007). For instance, instructors may include target formulaic sequences (Wood, 2009) in a text and then ask students to read it aloud repeatedly with increased time pressure over cycles, which may result in better retention of word chunks (Durrant & Schmitt, 2010). In addition, when using read-aloud or other scripted tasks, instructors can help students recognize where to pause and which words form a unit by using typographical enhancement, such as punctuation markers. However, the effect of such pedagogical interventions should be empirically examined in future research. When it comes to the use of read-aloud task in L2 fluency assessment, the current findings suggest that human raters (e.g., EIKEN) may be susceptible to the location of pauses (clause-internal vs. clause-external pauses), which should be reflected in the automated machine scoring in language tests (e.g., PTEA).

Although this study highlights how pause locations and articulation speed relate to perceived fluency during a read-aloud task, several factors may limit its generalizability. First, to minimize the influence of listeners' individual characteristics, we purposefully recruited L2 English-speaking raters who had L2 teaching experience but had little exposure to the Japanese language. Nonetheless, the raters had variation in their familiarity with Japanese-accented English (*M* = 27.3%, *SD* = 26.1). Although Kahng (2018) did not find any relationships between listeners' accent familiarity and L1 Korean speakers' fluency ratings, future research should explore if such relationships exist when different L1–L2 combination is concerned (e.g., listeners with varying degrees of familiarity with Japanese accents evaluate L2 English fluency). In addition, fluency in this study was operationalized by having the raters judge how smoothly the speech was delivered while focusing on temporal speech features. Although inter-rater reliability among raters was high, it would be important to qualitatively investigate which temporal measures of speech the raters focused on when evaluating fluency in a read-aloud task to triangulate the current findings. Finally, although the use of read-aloud tasks was ecologically valid for the Japanese EFL setting where there is little L2 exposure outside the classroom (Uchida & Sugimoto, 2018), future investigations of speech fluency should explore the relationship between utterance and perceived fluency for other tasks and in other foreign and second language contexts.

Notes

[1] Familiarity with L2 accent has often been measured based on numerical scales with end descriptors (e.g., 1 = *not at all*, 6 = *very much*: Magne et al., 2019). In this study, a percentage scale with end descriptors was used so that rater background variables (e.g., daily English use, accent familiarity) are numerically comparable, but as the Editor pointed out, this unusual metric may be subject to validity threats.

[2] As recommended by Suzuki et al. (2021), we used articulation rate rather than MLR because the latter reflects multiple dimensions of utterance fluency.

Acknowledgements

We gratefully acknowledge the great assistance of Shungo Suzuki in data coding and help with analyzing the data. This study was funded by a grant awarded to Kim McDonough from the Canada Research Chairs program (950-231218).

Aki Tsunemoto is a PhD candidate in Education at Concordia University. Her research interests include second language speech assessment and individual differences in speech perception.

Pakize Uludag is an Assistant Professor in technical communication in the Centre for Engineering in Society at Concordia University. Her interests include academic writing, language assessment and Corpus Linguistics.

Kim McDonough is a Professor of Applied Linguistics at Concordia University. Her research examines visual cues during task-based interaction, reverse linguistic stereotyping, and writing development.

Talia Isaacs is Associate Professor of Applied Linguistics and TESOL at the IOE, UCL's Faculty of Education and Society. Her research interests include speaking and assessment.

References

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

Boersma, D., & Weenink, P. (2017). *Praat: Doing phonetics by computer* (Version 6.0.40). http://www.praat.org

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

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation hardback with audio CDs: A Course book and reference guide*. Cambridge University Press.

Coulmas, F., & Watanabe, M. (2011). Japan's nascent multilingualism. In L. Wei, J. Dewaele & A. Housen (Ed.), *Opportunities and Challenges of Bilingualism* (pp. 249–272). De Gruyter Mouton.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners’ fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America*, *111*(6), 2862–2873. https://doi.org/10.1121/1.1471894

De Jong, N. H. (2018). Fluency in second language testing: insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237–254. https://doi.org/10.1080/15434303.2018.1477780

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech* (pp. 17–20). Royal Institute of Technology (KTH).

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

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research, 26*(2), 163–188. https://doi.org/10.1177/0267658309349431

EIKEN. (n.d.). 受験の状況 [Examinee statistics]. Retrieved February 22, 2022, from https://www.eiken.or.jp/eiken/merit/situation/

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics*. SAGE.

Hayes-Harb, R., Nicol, J., & Barker, J. (2010). Learning the phonological forms of new words: Effects of orthographic and auditory input. *Language and Speech*, *53*(3), 367–381. https://doi.org/10.1177/0023830910371460

Isaacs, T. (2018). Fully automated speaking assessment: Changes to proficiency testing and the role of pronunciation. In O. Kang, R. I. Thomson, & J. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 570–584). Routledge.

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing, 37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics, 39*(3), 569–591. https://doi.org/10.1017/S0142716417000534

Koizumi, R. (2022). L2 speaking assessment in secondary school classrooms in Japan. *Language Assessment Quarterly.* Advance online publication. http://doi.org/10.1080/15434303.2021.2023542

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

Laan, G. P. M. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, *22*(1), 43–65. https://doi.org/10.1016/S0167-6393(97)00012-5

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* Routledge.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387–417. http://doi.org/10.1111/j.1467-1770.1990.tb00669.x

Magne, V., Suzuki, S., Suzukida, Y., Ilkan, M., Tran, M., & 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

Pennycook, A. (2020). The future of Englishes: One, many or none? In A. Kirkpatrick (Ed.), *The Routledge handbook of world Englishes* (pp. 679–692). Routledge.

Plonsky, L., & Oswald, F. L. (2014), How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878–912. https://doi.org/10.1111/lang.12079

Prefontaine, 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. *The Canadian Modern Language Review*, *65*(3), 395–412. https://doi.org/10.3138/cmlr.65.3.395

Saito, K., Ilkan, M., Magne, V., Tran, M., & 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

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

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition, 42*(1), 143–167. http://doi.org/10.1017/S0272263119000421

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., Nakatsuhara, F., & Hunter, A. M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *The Modern Language Journal, 104*(1), 169–191. https://doi.org/10.1111/modl.12620

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.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics, 17*(1), 84–119. https://doi.org/10.1093/applin/17.1.84

Trofimovich, P., Kennedy, S., & Blanchet, J. (2017). Development of second language French oral skills in an instructed setting: A focus on speech ratings. *Canadian Journal of Applied Linguistics*, *20*(2), 32–50. https://doi.org/10.7202/1042675ar

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2021). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition.* Advance online publication. http://doi.org/10.1017/S0272263121000425

Uchida, Y., & Sugimoto, J. (2018). A survey of pronunciation instruction by Japanese teachers of English: Phonetic knowledge and teaching practice. *Journal of the Tokyo University of Marine Science and Technology*, *14*, 65–75. https://ci.nii.ac.jp/naid/120006402286/

Weinberger, S. (2015). *Speech Accent Archive*. Retrieved from http://accent.gmu.edu

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *Canadian Journal of Applied Linguistics*, *12*(1), 39–57. https://journals.lib.unb.ca/index.php/CJAL/article/view/19898

Yoshimura, Y., & MacWhinney, B. (2007). The effect of oral repetition on L2 speech fluency: An experimental tool and language tutor. *SLaTE-2007*, 25–28. https://psyling.talkbank.org/years/2007/fluency.pdf

Appendix A

Read-Aloud Passage from Speech Accent Archive (Weinberger, 2015)

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

Appendix B

Background Questionnaire for Raters

I. General background

1. Name: ______________________

2. Nationality: ____________

3. Gender: Male / Female / Other

4. Age: _____________ (years)

5. Birthplace (City, Province/State, Country): __________________________

6. Current Degree/ Major/ Year of study (if applicable): _________________

7. Last Degree you earned/ Major: ___________________________________

8. Is your hearing normal as far as you know? Yes / No

II. Language use

9. What do you consider to be your native language (from birth)? __________________

10. If English is your native language, which variety of English do you speak? ________

(e.g., Toronto, New York)

11. What language do you use as the major medium of communication now? ____________

12. Approximately what percent of the time do you speak English (as opposed to other languages) in your daily life?

(0% 10 20 30 40 50 60 70 80 90 100%)

13. Approximately what percent of the time do you listen to the English language media (as opposed to the media in other languages)?

(0% 10 20 30 40 50 60 70 80 90 100%)

14. Which languages can you speak other than English (if any)? ________________________

15. Of the languages you listed above, which would you say you are proficient in?

____

16. Which of your parents are first language English speakers? Mom / Dad / Both / Neither

17. Period of residence in English-speaking countries: ________________________________

18. If you were ever schooled in a language other than English as the primary medium of instruction, please specify which language in the table below. If English was the predominant language throughout your schooling, please skip to the next question.

Educational level	Language of instruction (if not English)
Primary	
Secondary	
Undergraduate	
Graduate	

III. Teaching experiences and Familiarity with Japanese

19. Do you have any teaching experience? Yes / No

20. If yes, please describe the context (place, year, length, subject):

____

21. Do you have any pronunciation training or taken a phonology course? Yes / No

22. If so, please describe the context:

____

23. Have you had any linguistics background (i.e., linguistics major, classes)? Yes / No

24. Have you taken any Japanese language related courses?   Yes / No

25. Approximately, how familiar are you with foreign accented English?

(0% 10 20 30 40 50 60 70 80 90 100%)

Not at all   Very familiar

25. Approximately, how familiar are you with Japanese accented English?

(0% 10 20 30 40 50 60 70 80 90 100%)

Not at all   Very familiar

26. Approximately, how often do you have contact with native Japanese speakers?

(0% 10 20 30 40 50 60 70 80 90 100%)

Very infrequent   Very often