Listener Response Time in Comprehensibility, Accentedness, and Fluency Judgments

Katherine Yaw

Northern Arizona University

Abstract

Listener perceptions of speakers' linguistic abilities contribute to successful communication in spoken interactions. The current pilot study investigated the amount of time listeners needed to make ratings of speakers' comprehensibility, accentedness and fluency. Speaker participants were 15 non-native speakers of English recruited from the Program in Intensive English (PIE) and the Graduate Program in Applied Linguistics at Northern Arizona University (NAU); listener participants were 13 native and non-native speakers of English recruited from first-year English composition and other courses at NAU. Listeners completed a survey in which they rated the 15 speakers on comprehensibility, accentedness, and fluency while their response times for each rating were measured. Correlational analyses did not indicate a statistically significant relationship between response time and construct ratings, though this sample size was small. Future directions include comparison of phonological analyses of speech samples with listener response times, as well as measuring listeners' "willingness to listen" to compare with response times and ratings.

Keywords: comprehensibility, accentedness, fluency, listener response time, speech perception, willingness to listen

Listener Response Time in Comprehensibility, Accentedness, and Fluency Judgments

Background

Studies in perception judgments of non-native speech have been used to inform second and foreign language instruction, particularly in pronunciation and other speaker interventions. Such studies include rating speakers on their comprehensibility (the ease with which a listener understands a speaker; see Lindemann & Subtirelu, 2013), accentedness (the degree to which a speaker's segmental and suprasegmental production varies from a listener's internalized, expected standard; see Munro, 2018), and/or fluency (perceived efficiency of a speaker's utterances and underlying cognitive processes; see Segalowitz, 2010). In the field of applied linguistics, however, there is growing interest in the role of the listener in ensuring successful communication. For instance, Kang, Rubin, and Lindemann (2015) have explored the use of intergroup contact exercises as a listener intervention with the aim of making listeners more receptive to different varieties of non-native speech, though the effects of such treatment are based on self-reported attitude data. While reporting judgments of accentedness, comprehensibility, and fluency is one way to measure a listener's attitudes towards a speaker, these measures only tell part of the story. Listener cognitive processes also contribute to these judgments and overall attitudes.

To date, only one study (Munro & Derwing, 1995) has attempted to investigate these cognitive processes in relation to listeners' perception judgments of speech. They found that speech rated as more accented took longer to process, and that listeners account for processing time in their determinations of how easily they can comprehend a speaker's message. These findings suggest that response time, a listener behavior that can be measured independent of

what a listener reports, reflects a level of difficulty or effort required in processing speech. Studies of response time in speech ratings, therefore, may find that when a listener chooses to make a rating reflects features of speech (phonological, lexical, grammatical, or pragmatic) that serve as "tipping points" in a listener's mental equation of a speaker's abilities. Additionally, response time may reflect a listener's "willingness to listen," or the degree to which a listener will stick with a speaker in a communicative interaction (Roberts & Vinson, 1998; Richmond & Hickson, 2001). While the current study does not focus on this final concept, it is an area that may merit future investigation when considering how to interpret response time data and translate such results into concrete listener intervention designs.

The current study sought to pilot a research design to extend Munro and Derwing's (1995) work by measuring how much time listeners take when making judgments of comprehensibility, accentedness, and fluency of non-native speech, both overall and by individual construct. In this case, response time provided an indication of when conscious judgments are made by the listener. The goal for this study design was to establish a baseline for the relationship between response time and construct ratings to inform future research and eventual practical applications.

Research Questions

The following research question and sub-questions were explored:

RQ1: What is the overall relationship between listener response time and listener ratings of speech samples?

RQ1a: What is the relationship between listener response time and comprehensibility ratings?

RQ1b: What is the relationship between listener response time and accentedness ratings?

RQ1c: What is the relationship between listener response time and fluency ratings?

Method

This pilot study included two phases of data collection: speech sample recording and speech stimulus rating. Similar to studies that aim to include speakers from a range of L1 backgrounds, this study employed a verbal guise design, meaning that each speaker contributed one speech sample to the pool of speech stimuli (Garrett, 2010).

Participants

Speaker participants in this study were volunteers recruited from the Program in Intensive English (PIE) and the graduate program in Applied Linguistics at Northern Arizona University.

Table 1

Speaker Participant Characteristics

Speaker Characteristic		*n*	%
Gender			
	Female	7	47%
	Male	8	53%
First Language			
	Arabic	7	47%
	Chinese	4	27%
	Farsi	1	6.6%
	Korean	1	6.6%
	Spanish	1	6.6%
	Vietnamese	1	6.6%
Proficiency			
	PIE Level 2 (iBT 16-32)	2	13%
	PIE Level 4 (iBT 44-56)	4	27%
	PIE Level 5 (iBT 57-69)	5	33%
	PIE Level 6 (iBT > 70)	2	13%
	Doctoral (iBT > 104)	2	13%

PIE students were offered pronunciation feedback in exchange for their participation in recording speaking tasks. A total of 15 speakers participated in this study. These speakers represented a range of L1s and proficiency levels, as is illustrated in Table 1.

Listener participants were recruited from two undergraduate first-year English courses at Northern Arizona University and were offered extra credit for participation in this study. Additional participants were recruited from the researcher's network of colleagues in TESOL and applied linguistics programs around the world.

Table 2

Listener Participant Characteristics

Listener Characteristic		*n*	%
Age			
	18-24	8	62%
	25-34	3	23%
	35-44	2	15%
Gender			
	Female	10	77%
	Male	3	23%
First Language			
	English only	11	85%
	Bilingual English/Other	1	8%
	Other	1	8%
Other Languages Spoken			
	Yes	5	38%
	No	8	62%
Other Languages Studied			
	Yes	9	69%
	No	4	31%
Linguistics Background			
	Yes	9	69%
	No	4	31%
Contact with NNS			
	At least once a day	5	38%
	At least once a week	3	23%
	At least once a month	2	15%
	At least once a year	3	23%
	Never	0	0%

A total of 13 listeners participated in this pilot (demographics shown in Table 2). As part of their participation, listeners were asked to provide information about their linguistic backgrounds, previous language study, and level of interaction with non-native speakers (NNS) of English. This information was gathered for future moderator variable analysis.

Speech Instrument and Recording Procedure

For the speech elicitation portion of this study, speakers completed a picture description task that asked them to describe the story of two people who collide while walking down the street and accidentally exchange the suitcases they were carrying (see Appendix A). Developed by Derwing, Munro, Thomson, & Rossiter (2009), this task was chosen because it prompts spontaneous language production which is suitable for maintaining novelty for ratings of comprehensibility. Speakers were given as much time as they needed to mentally prepare to complete the task, though they were not allowed to take notes or write out a script. They were encouraged to include as many details as they wanted and to be creative with their story.

All recordings were conducted in a quiet room with a Sennheiser SC 60 USB CTRL microphone headset and a MacBook Pro computer using PRAAT (Boersma & Weenink, 2018). Prior to each recording session, speakers completed a consent form and provided basic information about their language background (first language, plus any other languages they speak). They completed two additional tasks, an elicited imitation with six short sentences (Trofimovich, Lightbrown, Halter, & Song, 2009) and a read-aloud paragraph (Celce-Murcia, Brinton, Goodwin, & Griner, 2010), to familiarize them with the recording set-up and interacting with the researcher. These tasks were not used in the current study. Following the rating sessions, speech files were prepared for inclusion in the survey instrument. Each speaker's recording was cut into two 30-second clips each for rating in accordance with Kermad and

Kang's (2017) finding of no statistical difference between 30-second clips and longer clips in terms of rating for accentedness and comprehensibility. Having two clips per speaker, Clip A and Clip B, allowed for less repetition of speech stimuli in the rating phase.

Survey Instrument and Rating Procedure

The survey instrument used in this study was developed and administered via an online survey program, Qualtrics (2018). The survey included five sections: listener background questionnaire, comprehensibility rating block, accentedness rating block, fluency rating block, and qualitative reflection block. Listener background questionnaire items are included in Appendix B. The structure of the comprehensibility, accentedness, and fluency blocks were identical. Each block asked listeners to rate all 15 speakers. To mitigate order effects, each block included two levels of randomization: a) the order in which the speakers were presented, and b) whether the listener heard the A or B clip for each speaker. Listeners provided 15 ratings per construct, for a total of 45 ratings per listener. Qualitative items were three questions that asked participants to reflect on what speech features they felt they were listening for when rating each of the three constructs (see Appendix D).

For the speech rating portion of this study, participants were presented with 9-point semantic differential scales to rate comprehensibility (1 = very easy to understand, 9 = very difficult to understand), accentedness (1 = no foreign accent, 9 = very strong foreign accent), and fluency (1 = very fluent, 9 = not fluent at all). The use of a 9-point scale is consistent with Munro's (2018) findings that a scale of this type elicits ratings consistent with a direct magnitude estimation (DME) in which listeners heard a baseline stimulus and rated all later speech stimuli in comparison to the baseline stimulus. Listeners in the current study were instructed that they did not have to listen to a speaker's full speech file before making a rating; rather, they could

make their rating as soon as they felt ready. The rating instructions and an example of the question layout are included in Appendix C. Reliability of ratings for each construct was measured using Cronbach's alpha. Inter-rater reliability was highest for comprehensibility (α = 0.94), though accentedness (α = 0.86) and fluency (α = 0.78) exhibited reasonably high reliability for the size of the rater sample.

Results

Listener ratings and response time data were examined descriptively to check for normality assumptions. To do this, a mean rating and a mean listener response time were calculated for each of the 15 speakers. Tables 3 and 4 report the descriptive statistics for these 15 mean scores and response times. As can be seen in both tables, several of the skewness and kurtosis values fall outside of the range of -2 to 2, indicating that normality assumptions were not met.

Table 3

Descriptive Statistics for Mean Ratings by Construct

Measure	N	M	SD	Min	Max	Skewness	Kurtosis
Comp Rating	15	4.66	1.39	2.23	6.85	-0.15	-0.73
Accent Rating	15	6.03	1.32	2.54	7.54	-1.63	2.88
Fluency Rating	15	4.75	1.51	1.62	6.85	-0.56	-0.51
Combined Rating	15	5.14	1.33	2.13	6.97	-0.80	0.40

Table 4

Descriptive Statistics for Mean Response Times (RT) by Construct

Measure	N	M	SD	Min	Max	Skewness	Kurtosis
Comp RT	15	20.41	4.47	16.62	34.85	2.67	8.44
Accent RT	15	14.28	2.95	10.74	20.42	0.86	0.11
Fluency RT	15	13.42	2.09	8.20	17.56	-0.82	2.70
Combined RT	15	16.04	2.21	12.45	21.00	0.80	0.66

Given the lack of normal distribution among this data set, Spearman's rho correlation coefficients were calculated between the rating and response time data for each construct, as well as between the rating and response time data for all three constructs combined. Presented in Table 5, the results indicate significant and strong correlations among the different rating measures, and significant but moderate correlations among the response time measures. Neither of these results is surprising; these provide evidence of consistency in rating behaviors. What is unexpected is the moderate correlations between fluency ratings and both comprehensibility and combined response times that both reach significance at $p < 0.05$. This may be an artifact of the small sample size and is an area to watch as more data is added for these measures.

Table 5

Spearman's Rho Correlations among Ratings and Response Times (RT)

Measure	1	2	3	4	5	6	7
1. Comp rating							
2. Comp RT	**0.43**						
3. Accent rating	0.69**	0.22					
4. Accent RT	0.11	0.29	**-0.10**				
5. Fluency rating	0.95**	0.55*	0.77**	0.22			
6. Fluency RT	0.22	0.24	0.08	0.15	**0.23**		
7. Combined rating	0.96**	0.38	0.84**	0.09	0.96**	0.17	
8. Combined RT	0.41	0.76**	0.28	0.63*	0.55*	0.55*	**0.39**

Note. * $p < 0.05$, ** $p < 0.01$

Of interest in answering the research questions are the four bolded values in Table 5, which demonstrate the correlation between rating and response time for each construct, along with a correlation for the three constructs' ratings and response times combined. Comprehensibility displayed the highest value with a moderate correlation ($r = 0.43$, $p = 0.11$), followed by the combined constructs ($r = 0.39$, $p = 0.15$) and fluency ($r = 0.23$, $p = 0.40$). Accentedness presented with a weak negative correlation ($r = -0.10$, $p = 0.74$), indicating that when listeners rated speakers as less accented, it took them longer to make that rating. Notable

in all of these correlations is that none of them reached significance at $\alpha = 0.05$. This, again, may be a reflection of the small sample size, though it may also reflect the instrument used to measure response time.

Discussion

Based on the correlation data presented in the previous section, there is no significant relationship between listener construct ratings and the time it takes them to make these ratings. With a listener sample of only 13 participants, any correlational values should be interpreted cautiously, as *p* values in small samples are less stable. However, there is another important consideration for understanding this data: the response time measure.

In Qualtrics, response time to a question is measured through the timing of clicks. Qualtrics can record the time of the first click on a page, last click on a page, and click to submit a page, as well as the total number of clicks. The survey was designed with the assumption that the first click would start the speech file playing and the last click would represent the final rating decision, so that response time could be calculated as the difference between the two click times. When piloting the survey before administering it to participants, though, managing the click data was tricky. Qualtrics would not recognize a click on a play button of a speech file as a first click (it did not register any clicks on the speech file as clicks), making it necessary to find another way to control the start time for each question. The solution to this was to set each speech file to auto-play as soon as the question loaded. With this set-up, response time was then measured as the time of the last click (which was again assumed to be the final rating decision). From examining individual response times, it is clear that having a speech file play automatically did not always lead to the listener paying attention and being ready to listen. This, combined

with listeners completing the survey in a non-laboratory setting, leads to questions of how valid the existing Qualtrics instrument is at measuring the target cognitive process.

Additionally, in terms of the slightly negative correlation found between accentedness ratings and response times, qualitative data may provide some insight. One listener, when asked what they focused on when rating accentedness, commented, “This was much faster/easier to answer. You can almost immediately tell if someone has a strong accent, and it's a more polarized answer. Either they have an accent or they don't really have much of one. The ones who were considered to have no or very little accents usually required me to listen for a few more seconds than the ones with ‘strong’ accents. Strong accents are almost immediately identifiable to me.” This comment was interesting because the researcher’s expectation was that accents on either extreme of the spectrum would be rated faster than those deemed to be more in the middle, whereas this listener states that stronger accents elicited the fastest responses.

Piloting this study design revealed a number of limitations that need to be considered for future iterations of this study. The sample for this pilot was small at only 13 listener participants, and visual inspection of the data from this sample indicated that at least two of the participants completed the survey in a cursory manner (with an average response time of 2 seconds per rating, compared to the rest, who ranged from 10 to 30 seconds per rating). While this sample allows for preliminary data analysis, it is not large enough to provide for generalizable conclusions.

From a technological standpoint, there were numerous challenges in the design and implementation of this survey. Using Qualtrics to measure response time required setting the audio files to auto-play, which led to issues of accurate response time measurement as some listeners seemed to leave the question window open well beyond the time needed to make a

rating. Additionally, some listeners reported that they were unable to hear any of the audio files and were thus unable to participate. Finding a data collection process that removes these issues is a priority for this study moving forward. This may involve conducting survey sessions in a laboratory setting, and/or require the use of a different response time measurement tool (such as Paradigm).

Finally, the data analysis conducted with this project does not reflect the full range of what can be analyzed from a study with this design. Although this paper reports descriptive data and correlations between construct ratings and construct response times, there is opportunity for more nuanced analysis of mean differences within constructs by grouping speaker data according to proficiency in each construct, as well as analysis of individual listeners' and speakers' performance with this instrument. An additional level of analysis that is planned for this data will come from phonological analysis of segmental (number of errors) and suprasegmental (speech rate and pausing) features. Inclusion of phonological analyses allows for multiple regression analysis to determine which speech features contribute to comprehensibility, accentedness, and fluency ratings.

Relevance to the PIE and Second Language Learning

This pilot offers an informative first step in research on response time in studies of perception of speech. This is relevant to second language learning in that it provides a greater understanding of the behavior of listeners who interact with language learners, including PIE students. In addition to the limitations commented upon earlier, this pilot has identified some promising areas in the development of listener interventions, namely exploration of the "willingness to listen" construct in relation to listener response time and ratings.

References

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Version 6.0.39, retrieved 3 April 2018 from http://www.praat.org/

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

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*, 533-557.

Fuertes, J.N., Gottdiener, W.H., Martin, H., Gilbert, T.C., & Giles, H. (2012). A meta-analysis of the effects of speakers' accents on interpersonal evaluations. *European Journal of Social Psychology, 42*, 120-133. doi: 10.1002/ejsp.862

Garrett, P. (2010). *Attitudes to language.* Cambridge, UK: Cambridge University Press.

Jegerski, J. & VanPatten, B. (Eds.). (2014). *Research methods in second language psycholinguistics*. New York, NY: Routledge.

Kang, O., Rubin, D., & Lindemann, S. (2015). Mitigating U.S. undergraduates' attitudes toward international teaching assistants. *TESOL Quarterly, 49*, 681-706. doi: 10.1002/tesq.192

Kermad, A., & Kang, O. (2017, September). *Listener perception of pronunciation and length of speech stimuli: Does length matter?* Paper presented at the 9th annual conference of Pronunciation in Second Language Learning & Teaching, Salt Lake City, Utah.

Lindemann, S., & Campbell, M.-A. (2018). Attitudes towards non-native pronunciation. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 399-412). London, UK: Routledge.

Lindemann, S., & Subtirelu, N. (2013). Reliably biased: The role of listener expectation in the perception of second language speech. *Language Learning, 63*, 567-594. doi: 10.1111/lang.12014

Munro, M. J. (2008). Foreign accent and speech intelligibility. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and second language acquisition* (pp. 193-218). Amsterdam, Netherlands: John Benjamins.

Munro, M. J. (2018). Dimensions of pronunciation. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 413-431). London, UK: Routledge.

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech, 38*, 289-306.

O'Brien, M.G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning, 64*, 715-748. doi: 10.1111/lang.12082

Pinget, A.-F., Bosker, H. R., Quené, H., & de Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing, 31*, 349–365. doi: 10.1177/0265532214526177

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*, 53-73. doi: 10.1177/0265532215579530

Qualtrics (Version May 2018). [Survey software]. Retrieved from https://www.qualtrics.com

Richmond, V. P., & Hickson, M. III. (2001). *Going public: A practical guide to public talk.* Boston, MA: Allyn & Bacon.

Roberts, C. V., & Vinson, L. (1998). Relationship among willingness to listen, receiver apprehension, communication apprehension, communication competence, and dogmatism. *International Journal of Listening, 12*(1), 40-56.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review/La revue canadienne des langues vivantes, 65*, 395-412. doi: 10.3138/cmlr.65.3.395

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical correlates of comprehensibility versus accentedness in second language speech. *Bilingualism: Language and Cognition, 19*, 597-609. doi: 10.1017/S1366728915000255

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York, NY: Routledge.

Subtirelu, N., & Lindemann, S. (2016). Teaching first language speakers to communicate across linguistic difference: Addressing attitudes, comprehension, and strategies. *Applied Linguistics, 37*, 765-783. doi: 10.1093/applin/amu068

Trofimovich, P., Lightbrown, P. M., Halter, R., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition, 31*, 609-639. doi: 10.1017/S0272263109990040

Appendix A

Picture Description Task

Instructions: Look at the pictures below and think about the story that they show. When you are ready, please tell the story with as many details as possible. You can be creative with your story.

(Picture task taken from Derwing et al., 2009)

Appendix B

Listener Background Questionnaire

The following questions will ask for information about your background and language experiences.

B1 How old are you?

▼ 18 (1) ... over 90 (74)

B2 Which of the following do you identify as?

O Female (1)

O Male (2)

O Prefer not to answer (3)

B3 What is your first or native language?

O English (1)

O Other (please list) (2) _____

B4 What other language(s) do you speak?

O None (1)

O Other (please list) (2) _____

B5 What other language(s) have you studied?

- O None (1)
- O Other (please list) (2) _____

B6 Have you taken courses in linguistics, languages, and/or language education?

- O Yes (1)
- O No (2)

Display This Question:

If BLing = Yes

B7 Please describe your previous experience with linguistics, languages, and/or language education.

_____

B8 How often do you interact with non-native speakers of English?

- O At least once a day (1)
- O At least once a week (2)
- O At least once a month (3)
- O At least once a year (4)
- O Never (5)

Appendix C

Survey Instructions and Sample Rating Question

Survey Instructions: In this survey, you will listen to 15 different speakers telling parts of a story about the following picture. It is recommended to use headphones for the rest of this survey.

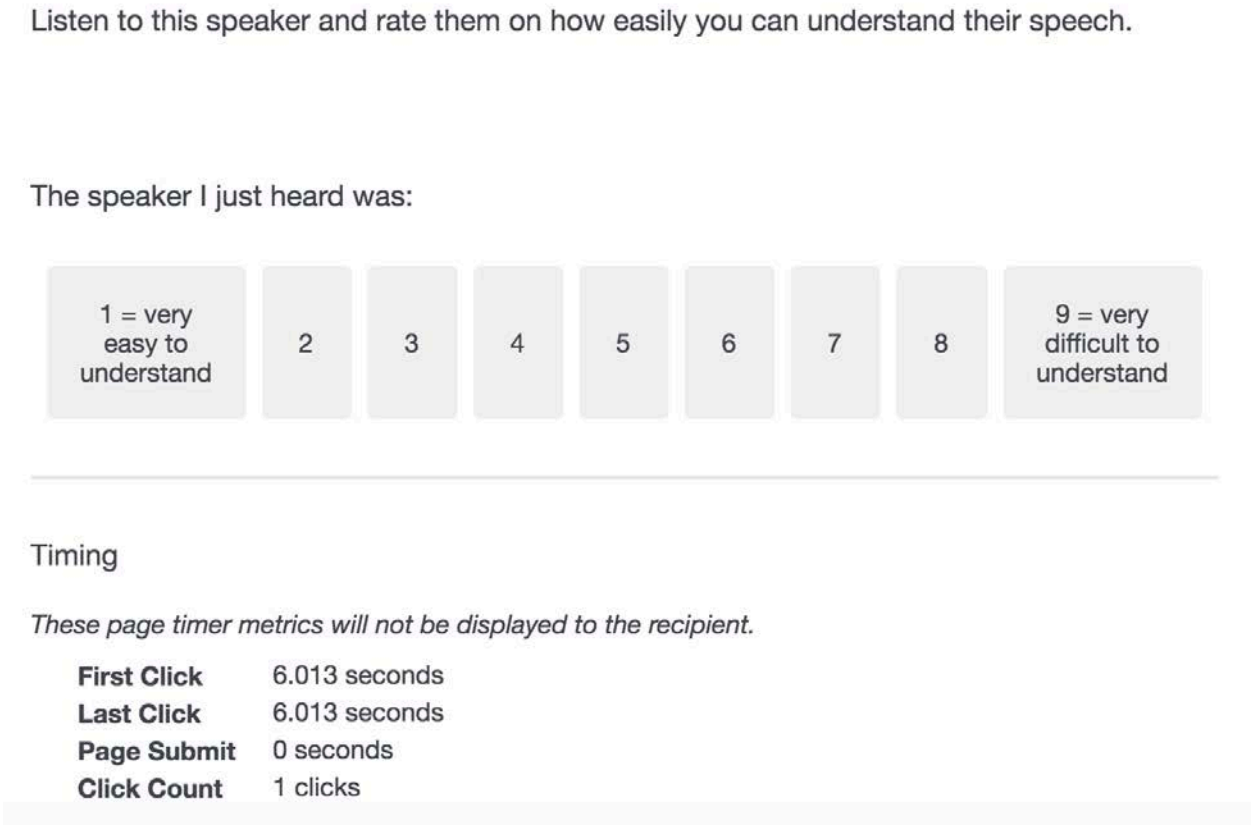
THE SUITCASE STORY

You will be rating each speaker on their comprehensibility, accentedness, or fluency. As soon as you are ready to make your rating, please click your score (1-9) on the scale. You can make your rating at any point; you do not have to wait until the speaker finishes talking.

IMPORTANT: The sound files will automatically play when you move to a new question. Each speaker talks for 30 seconds, though you do not need to listen to the whole file if you are ready to make your rating earlier. If you need to take a break, you can do this at the end of each section (it will be marked on the screen).

Next you will practice rating two examples that represent the range of speakers you will hear. Please have your headphones on and volume ready before you click "next."

Rating Item Layout: Listeners only saw the instructions and rating scale on their screen. The audio file automatically played. The timing information below the rating scale was recorded by Qualtrics, but it was not visible to the listeners while completing their rating tasks.



Appendix D

Survey Qualitative Items

You have completed the listening portion of this survey. In the questions below, please provide some feedback on what you were listening for while rating these speakers.

QualComp What features of the speech samples do you think impacted your comprehensibility ratings the most? In other words, what did you pay attention to most when deciding how comprehensible a speaker was? Why?

_____

_____

_____

_____

_____

QualAcc What features of the speech samples do you think impacted your accentedness ratings the most? In other words, what did you pay attention to most when deciding how much of a foreign accent a speaker had? Why?

_____

_____

_____

_____

_____

QualFlu What features of the speech samples do you think impacted your fluency ratings the most? In other words, what did you pay attention to most when deciding how fluent a speaker was? Why?

_____

_____

_____

_____

_____

QualOtherComments If you have any other comments regarding this survey or the rating process, you are welcome to share them here.

_____

_____

_____

_____

_____