Exploring Paused Transcription to Assess L2 Listening Comprehension Utilizing Rasch Measurement

Allie Patterson
patterson.allie@nihon-u.ac.jp
Nihon University School of Medicine

Abstract

In-class L2 researchers often do not have large research budgets and do not have access to brain imaging technology. Access to funds and this technology is usually required to explore L2 listening processing in a meaningful way. A relatively new method developed by Field (2008) and further refined by Yeldham (2016) called paused transcription shows promise as a cheap method for testing L2 listening but has not been analyzed with an eye towards validity until now. In this study, a paused transcription listening test was developed for use in a mixed effects model (LME) study to be conducted at a future date. This instrument was administered to 37 first year Japanese students. A Rasch analysis showed that the instrument had high item and student reliability. Dependence between items was also found but is expected in this type of method and can be controlled for in future analyses.

Keywords: Listening, transcription, functor, content, EFL

Listening in an L2 is a cognitively taxing activity. When processing L2 speech, it is likely a listener prioritizes some forms over others. Word attributes, such as stress and frequency in speech, likely have an impact on whether a heard word is successfully processed by a L2 listener. A handful of studies have broached this subject. However, the effects of word attributes on L2 listening processing are an underexplored aspect of SLA. In this study, I examine a method known as paused transcription which shows promise as a cost-effective means of researching L2 listening. This method was first developed by Pemberton (2004, as cited by Field, 2008) and later refined by Field (2008). This method has also been used in a study by Yeldham (2016).

Despite the proliferation of this method in recent research, no prior studies have field tested their instruments prior to conducting a study or attempted to create a validity argument for the method. This lack of validation for paused transcription calls into question the results of prior studies. In this brief study, I begin to fill this gap by conducting a field test of a self-developed paused transcription test. The results of this field test are then subsequently analyzed using Rasch analysis.

The rationale for studying differences in comprehension rates of heard speech based on word attributes comes from psycholinguistic research. The effects a variety of word attributes have on processing of language have been explored in previous research which has shown that the presence or absence of some word attributes can cause it to be prioritized over other words in processing. Prior psycholinguistic research has shown that categories such as functors and content words are processed in different parts of the brain (Brown, Hagoort, & ter Keurs, 1999). Low frequency words have been shown to cause more brain activity (Hauk & Pulvermuller, 2004) and larger pupillary dilation in subjects (Kuchinke, Vo, Hofmann, and Jacobs, 2007). Psycholinguistic research has also shown that longer words produce more brain activity in participants (Pulvermuller, 2004).

SLA research on the differences word attributes cause in processing usually relies on very different methods than those utilized in psycholinguistic research. The psycholinguistic studies cited previously rely on brain imaging or eye-tracking technology. However, SLA research is often classroom based and SLA researchers often do not have access to these types of technology. As such, SLA studies often rely on recall and comprehension tests to ascertain differences in processing caused by word attributes. For instance, Hahn (2004) tested the effect of misplaced stress on comprehension using a comprehension test.

Graham and Santos (2013) also used a recall test to find differences in successful processing of nouns and verbs. While this research provides valuable insights into the processing of heard L2 speech, recall tests usually occur long after the speech has been processed. They cannot make instantaneous measures of listener's responses to speech the way brain imaging or eye tracking technology can. Top-down processing and the limits of learner memory likely affect the results of studies that use delayed recall methods and comprehension tests.

Paused transcription allows SLA researchers to get more immediate feedback on how L2 listening is processed. Transcription in SLA research is most often associated with speaking research (Ellis, 2008). Transcription is often found in the methods employed in phonology research and conversations analysis where the researcher transcribes segments of speech produced by a L2 speaker. However, in recent years, some researchers have turned the tables and had students transcribe portions of heard L2 speech. Most SLA educators and researchers would associate having students do transcription with outdated teaching methods such as grammar translation. However, research conducted thus far using transcription has provided some interesting insights. In the paused transcription method, students are played either a monologue or a dialogue. Within the test audio file, pauses are intermittently inserted before target phrases. Participants are instructed to immediately transcribe the words preceding the pause. The transcribed phrases are then analyzed according to word attributes with each word acting as a separate item.

The handful of studies that have used this methodology have also been primarily concerned with differences in processing of content words and function words (functors). The first study to use student transcription as its principal method of investigation was Pemberton (2004, as cited by Field, 2008). He investigated differences in the percentage of uptake and successful processing between content words and functors. He did not find significant differences between transcribed amounts of functors and content words. However, Field (2008) argues that Pemberton's methods are problematic because participants were allowed to rewind recordings and listen to target phrases multiple times. While repetition does occur intermittently in speech (Rost, 2016), listeners are not actually able to have speakers repeat sentences verbatim repeatedly. As such, paused transcription as employed by Pemberton was artificial and is not a sufficient proxy for actual L2 listening processing.

Field (2008) expanded upon and improved the paused transcription method to better reflect the reality of L2 listening. In this study, L2 students listened to a recording of an interview. The recording was played within the student's classroom and was administered to all participants at once. Pauses were inserted into a recording of an interview and participants were instructed beforehand to write the last four or five words they heard before the pauses. Misspellings that phonetically approximated the target words were counted as correct instances of transcription. Field found that there were statistically significant differences between the number of content words and functors transcribed with content words being transcribed at a higher rate than functors. Unlike Pemberton's study, Field did not allow participants to rewind the recording. Field's version of the paused transcription methodology also had the benefit of being easily administered at once to an entire L2 classroom. The only materials needed to administer this type of test is a recording with pauses inserted after target phrases and a test form for participants to write their answers.

A recent study by Yeldham (2016) further improved the paused transcription method and was the first to recognize that paused transcription could be used to analyze the effects of word characteristics beyond differences between functors and content words. This study included the analysis of content words and functors found in the previous studies and found similar statistically significant differences with functors being transcribed at a lower rate. Yeldham also improved upon the method by including gaps in participant comprehension in the analysis. Field (2008) excluded student responses from analysis if none of the target words were transcribed in a phrase, but Yeldham (2016) included these blank instances in analyses to

better reflect the nature of L2 listening processing which would likely have large gaps in successful processing and comprehension. In addition to doing a functor/content word analysis, Yeldham included an additional analysis of the differences in transcription rates between stressed and unstressed functors. Yeldham found that stressed functors were being transcribed at a higher rate and hypothesized that more attentional resources were being devoted to these forms. This additional analysis shows that the paused transcription method could possibly be utilized to explore the effects a myriad of different word attributes have on successful processing of L2 heard speech.

For SLA researchers, paused transcription represents a viable alternative to recall tests, comprehension tests, and brain imaging studies. With paused transcription, participants are tasked with immediately writing what they just heard a few seconds prior. The immediate nature of the method helps control for top-down processing and does not depend on participant's long-term memory ability unlike recall and comprehension tests. Also, unlike brain imagining studies, the tools to conduct this research are readily available in L2 classrooms. All a teacher requires to conduct this type of research is an audio system, a recording, and test forms. L2 teachers do not require extensive knowledge of brain imaging technology and large research budgets.

Research Questions

While paused transcription shows promise, none of the prior research that utilizes it has questioned the validity of the method. Field (2008) and Yeldham (2016) do not include any mention of field testing their instruments prior to the study or conducting additional analyses to verify that data they are receiving from students is indeed representative of how listeners are processing heard L2 speech. This gap is problematic and hinders the generalizability of these studies. It is possible that this method is not a measure of successfully processed words but is actually a measure of some other phenomenon. In this study, I will attempt to rectify this lack of verifying evidence by beginning to construct a validity argument through the use of Rasch analysis. According to Fulcher and Davidson (2007), a validity argument is "the defense of a claim, requiring grounds (data) to support the claim, and a warrant to justify the claim on the basis of the grounds" (p. 377). The claim that I am hoping to defend in this study is that paused transcription is a valid format of analysis that can provide useful insights into the nature of L2 listening.

In order to defend this claim, this study will include four research questions. The first research question is concerned with the relationship between content words and functors. As stated, prior research using paused transcription has shown that content words are transcribed at statistically significant higher rates than function words (Field, 2008; Yeldham, 2016). One means of arguing for the validity of this particular test is to see if it is eliciting behavior that is similar to other assessments used in past research. As such, difference in transcription rates between content words and functors will tested. Research questions two and three are concerned with how well this assessment meets the assumptions of Rasch analysis. The final research question is concerned with the identical nature of several of the items. Due to the grammatical necessities of English, some words (i.e., *the* and *a*) are used several times in the instrument. It stands to reason that these items may show dependency, which could prove problematic in future analyses using data generated by this instrument. The research questions are as follows:

1. Are function words substantively and statistically significantly more difficult than content words?
2. Did the dataset show acceptable fit to the Rasch model?
3. Did the dataset show acceptable unidimensionality?
4. Was item dependency between items testing the same word observed?

Method

Instrument

I constructed a paused transcription test with the intent of using it in a study which utilizes a mixed effects model (LME) analysis to parse out which word characteristics have the largest effects on successful comprehension of heard L2 speech. A mixed effects model analysis allows researchers to test for nested random and hierarchical effects in data (Cunnings & Finlayson, 2015). In essence, a large number of fixed independent variables and random effects can be accounted for and their effects on the dependent variable will be quantified. The independent variables that will be accounted for in the future study will be word attributes such as word length, frequency in speech, stress, and imageability. This mixed effects model study will be conducted at a future date. This current pilot study was conducted in order facilitate this future study. The test specifications used to create this instrument can be observed in Appendix A.

The instrument is a recording of a monologue of an L2 teacher informing students about upcoming assignments. The full monologue with target phrases underlined can be observed in Appendix B. This subject matter of a language teacher talking about upcoming assignments was chosen because it was believed all L2 students would have the necessary experience to understand the content due to having extensive time operating in an EFL classroom. The instrument was created using a digital voice recorder and the audio software Audacity® version 2.2.2 (audacityteam.org). I used my own voice for the recording. For the test, the recording was embedded in a PowerPoint presentation. The PowerPoint also included a practice phrase which was used to model test procedures to the students before conducting the test.

An assortment of words with very different attributes was included in the instrument to create variation for the mixed effects analysis. Words in the target phrases were chosen with the aid of the MRC Psycholinguistic Database (2018). This database provides lists of words attribute variables that may affect processing, such as imageability and frequency. The target phrases can be observed in Table 1. Digital beeps and fifteen seconds of silence were digitally inserted into the recording after target phrases.

Table 1
Test Target Phrases

1. of work to do soon
2. Next week send it through
3. The subject of the essay
4. about your mother and father
5. What is their personality like
6. the city where you live
7. will have the grammar test
8. I can help you with
9. day is a national holiday
10. if there is a question
11. my desk before you go
12. guys got very good grades

For the field test of the instrument, the instructions were provided in English. However, in the actual study that will be analyzed with the mixed effects model, the instructions for the test will be provided in the students' L1, Japanese. The test instructions were included in the PowerPoint and on the test form. Below

the instructions on the test form, students were provided with twelve blank spaces where target phrases could be written. The test form can be observed in Appendix C.

Participants

The field test was administered to 37 first year Japanese students at a private university in Tokyo. The instrument was administered to a mixture of men and women during a TOEFL prep course.

Procedure

Permission to conduct this research was given by the management of the English program at the university. The test was not administered by me but by another EFL instructor. I was not present for the administration. This instructor was trained on the test procedures prior to administering the test. The instructor was also provided with test administration instructions that can be seen in Appendix D. Students were informed they would participate in a listening activity for research purposes where they would be tasked with quickly writing spoken English they heard.

Data Coding and Analysis

The analysis used in this study was Rasch analysis (Bond & Fox, 2015). The model used in this study is the original Rasch model developed for analysis of dichotomous data. Winsteps (Linacre, 2019), a type of software developed for conducting various forms of Rasch analysis, was used in this study. Through Rasch analysis, it is possible to see if any test items are eliciting odd behavior from test takers and how reliable the test is from student to student.

For the analysis, each word was treated as a separate item (n =60). The results of the test were first coded directly on the test sheet. In keeping with the methods used in Field (2008) and Yeldham (2016), words that phonetically approximated the target word, but had misspelling errors were counted as correct items. A *1* was given for correct transcription of a word and a *0* was given when the word was missing or when what was written did not phonetically approximate the target word. For example, an answer of *werk do soon* for the first target phrase with the first and third word missing and a spelling mistake on the second word would be coded as 01011. All answers were first coded, and these codes were then transferred to a Winsteps command file. Phrases with no target words transcribed by students were still included in the analysis. Multiple instances of words across target phrases were given a number corresponding with their appearance in the Rasch analysis to aid in distinguishing them. For instance, the first appearance of *the* is called *the1* and the second instance is *the2*. Students are identified through a four-digit number. Each item was also given a label to indicate if it was a content (*C_*) or functor (*F_*) word to facilitate the comparison necessary to answer the first research question.

Results and Discussion

Table 2 shows the overall reliability, separation, and fit statistics for the instrument. The test had a high overall participant reliability and a high item reliability. Figure 1 is a Wright map displaying the relationship between item difficulty and student ability. Students higher on the scale have higher ability, and items higher on the scale are more difficult. The results of the test as displayed by the Wright map are in line with the results found in Field (2008) and Yeldham (2016). Easier items, for the most part, are content words. For example, the easiest items on the test seen at the bottom of the Wright map were *father*, *mother*, *national*, and *holiday*. Three of the four most difficult words (*their, about, and soon*) were function words. A t-test was conducted between content and function words to answer the first research question. The results of the t-test were significant ($t(57) = -2.94$, $p = .005$) with functors being -1.23 logits more difficult than content words.

Table 2
Reliability, Separation, and Fit Statistics

	Reliability	Separation	Infit MNSQ (Min, Max)	Infit ZSTD (Min, Max)	Outfit MNSQ (Min, Max)	Outfit ZSTD (Min, Max)
Student	.92	3.5	(0.6, 1.6)	(-2.7, 2.5)	(0.3, 2.6)	(-2.2, 1.5)
Item	.92	3.3	(-1.5, 2.4)	(-1.5, 2.4)	(0.2, 4.3)	(-1.3. 3.0)

The Wright map shows that students are somewhat evenly distributed in terms of ability as measured by this test. There are no large groups of students at either end of the Wright map. There is a small group of students that performed better than their peers at the top of the Wright map. These are likely students who have studied or lived abroad who are highly proficient. The lack of large groupings at either end of the scales indicates this test is not too difficult or too easy for the majority of the participants. This spread is a desirable result because this test will be used in further research. Differences in outcome is desirable to ensure variation for statistical analyses. However, if adopted for classroom use, this activity would likely be used for criterion reference purposes. Teachers hoping to use paused transcription for classroom activities should ensure target phrases are easier than those adopted for the test in this study to ensure most students can successfully complete the activity.

Next, to answer the second research question, the infit and outfit statistics of items and participants were checked. Infit and outfit quantifies an item or student's adherence to the expectations of the Rasch model (Bond & Fox, 2015). Infit and outfit statistics help test administrators judge if test items and participants are exhibiting odd behavior or if items are unreliable measures of student ability. Items and participants were judged to fit the Rasch model if they fit within the range of .5 to 1.7, which Wright and Linacre (1994) judge to be acceptable for clinical observation.

Table 3 shows item infit and outfit statistics. Only one item, *A1*, (the first instance of the word *a*) is misfitting. This item is misfitting due to one high ability student missing the item and several low proficiency students answering it correctly. This item preceded two of the easiest items on the test, *national* and *holiday*. It is possible that this low salience article directly preceding two highly salient content words caused this odd behavior. While this type of behavior would be problematic for most types of tests, observing this type of interaction may provide L2 researchers with deeper insight into the true nature of L2 processing. It is possible that highly salient content word phrases such as *national holiday* monopolize a L2 listener's attention and cause less salient forms like articles to be dropped more often than if they were followed by a less salient content word. This hypothesis is only speculative and requires further research.

Several of the items found at the bottom of the table in Table 3 (*where, if, is2,* and *their*) are overfitting. According to Bond and Fox (2015), overfit is when an item adheres too closely to the Rasch model and does not display enough variation. Items that are overfitting are thought to be muted and overfitting is often due to dependency. Dependency is when performance on one item affects another item. Dependency can be problematic on tests that have items which are meant to be independent, such as multiple-choice tests. Dependency will be discussed further when Research Question 4 is addressed.

Table 4 shows student infit statistics. Two students, 1001 and 1032, were misfitting. A look at their responses shows that these students successfully transcribed a few high difficulty words but did not perform well overall on the test. Student 1032 did not transcribe any words until the test was almost complete. This may show that this student did not understand the test procedures until the test was almost finished. Conducting the test instructions in the student's L1 in the actual study may prevent this type of confusion. It appears that student 1001 did not remain on task, because they almost successfully transcribed two complete phrases at the beginning and end of the test but left other phrases blank. It is

difficult to control for this type of behavior on a test using a cognitively demanding method such as paused transcription. Despite these outliers, most of the participants had acceptable fit statistics.

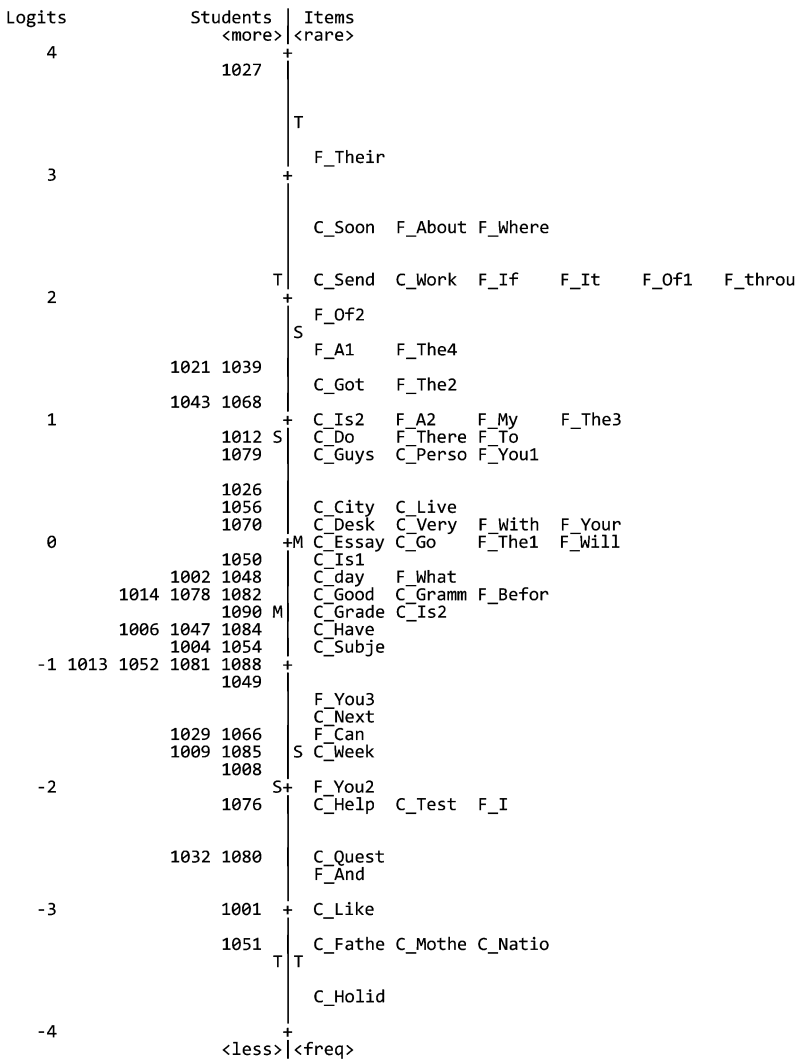

Figure 1. Wright map of relationship between item difficulty and participant ability.

Table 3
Item Infit and Outfit Statistics

	Infit		Outfit			
Items	MNSQ	ZSTD	MNSQ	ZSTD	Percent Correct	Point-measure Correlation
F_A1	1.7	1.8	4.3	3.0	16.2	.04
F_My	1.2	0.9	1.9	1.6	21.6	.28
F_Your	1.0	0.2	1.9	2.1	35.1	.43
C_Like	0.9	-0.2	1.7	1.0	86.5	.34
C_Desk	1.4	1.9	1.6	1.6	35.1	.24
F_What	1.4	2.4	1.5	1.4	43.2	.22
C_day	1.1	0.6	1.4	1.3	43.2	.38
F_The1	1.1	0.5	1.4	1.1	37.8	.41
C_Is1	1.3	1.5	1.3	1.0	40.5	.33
C_Subject	1.1	0.7	1.3	0.9	54.1	.38
F_Before	1.2	1.4	1.2	0.8	45.9	.35
C_Go	1.1	0.5	1.2	0.6	37.8	.43
C_Soon	0.9	0.1	1.2	0.5	8.1	.40
C_Is2	1.1	0.4	1.2	0.5	48.7	.44
C_Question	1.1	0.6	0.9	0.1	81.0	.31
C_Test	1.1	0.6	0.9	0.1	75.7	.35
F_And	1.1	0.4	0.9	0.2	83.8	.30
C_Work	1.1	0.4	0.9	0.2	10.8	.38
C_Grammar	1.1	0.6	1.0	0.2	45.9	.44
C_Very	1.1	0.6	0.9	-0.2	35.1	.46
F_With	1.1	0.5	1.0	0.2	35.1	.44
C_Week	1.1	0.4	0.9	0.0	70.3	.42
C_Personality	1.0	0.2	1.1	0.3	27.0	.45
F_A2	1.0	0.2	1.1	0.3	21.6	.45
F_To	1.0	0.2	0.9	0.1	24.3	.47
C_Do	1.0	0.2	0.9	0.1	24.3	.47
F_About	0.9	0.0	0.9	0.4	8.1	.42
C_Mother	0.9	0.0	0.8	0.0	89.2	.34
F_You3	0.9	-0.2	0.9	-0.1	62.2	.49
C_Father	0.9	0.0	0.7	-0.1	89.2	.35
C_Have	0.9	-0.4	0.9	-0.3	51.4	.52
C_Help	0.9	-0.3	0.7	-0.3	75.7	.47
F_The3	0.9	-0.3	0.8	-0.3	21.6	.52
C_National	0.9	-0.2	0.6	-0.3	89.2	.39
C_Essay	0.9	-0.7	0.7	-0.7	37.8	.57
C_Next	0.9	-0.8	0.7	-0.6	64.9	.56
F_Of2	0.9	-0.3	0.5	-0.5	13.5	.55
F_The2	0.9	-0.4	0.6	-0.6	18.9	.56
F_Can	0.9	-0.9	0.7	-0.5	67.6	.55
C_Send	0.8	-0.3	0.6	-0.2	10.8	.52

Table 3 (continued)

C_City	0.8	-0.9	0.7	-0.9	32.4	.60
C_Holiday	0.8	-0.3	0.3	-0.6	91.9	.42
C_Got	0.8	-0.6	0.8	-0.2	18.9	.55
F_The4	0.8	-0.6	0.6	-0.5	16.2	.58
F_Of1	0.8	-0.4	0.4	-0.6	10.8	.58
F_You1	0.8	-1.1	0.6	-1.0	27.0	.63
F_There	0.7	-1.1	0.5	-1.1	24.3	.64
F_I	0.7	-1.5	0.5	-0.8	75.7	.60
F_Where	0.6	-0.6	0.2	-0.6	8.1	.61
F_If	0.6	-0.8	0.3	-0.9	10.8	.64
F_Is2	0.6	-1.5	0.4	-1.3	21.6	.69
F_Their	0.6	-0.4	0.2	-0.4	5.4	.56

Table 4
Student Infit and Outfit Statistics

	Infit		Outfit			
Student	MNSQ	ZSTD	MNSQ	ZSTD	Percent Correct	Point Measure Correlation
1001	1.4	1.5	2.6	1.5	11.7	.22
1032	1.5	1.8	1.7	1.0	15.0	.28
1012	1.0	0.5	1.6	1.5	63.3	.48
1052	1.2	0.9	1.6	1.5	33.3	.50
1009	1.6	2.2	1.5	1.0	23.3	.35
1054	1.5	2.5	1.5	1.3	35.0	.38
1047	1.2	1.2	1.3	1.1	36.7	.50
1088	0.9	-0.6	1.3	0.8	33.3	.62
1039	1.3	1.5	1.1	0.4	71.7	.38
1027	1.2	0.5	1.3	0.6	95.0	.12
1014	1.1	0.4	1.2	0.7	41.7	.56
1080	1.2	0.7	0.8	-0.1	15.0	.46
1006	1.2	0.9	1.2	0.6	38.3	.53
1090	1.2	1.0	1.1	0.5	40.0	.53
1043	1.1	0.9	0.9	-0.1	66.7	.48
1068	1.0	0.5	0.9	-0.2	66.7	.51
1048	1.0	0.3	0.9	-0.2	45.0	.58
1026	1.0	0.2	0.9	-0.2	55.0	.57
1021	1.0	0.0	0.8	-0.2	71.7	.50
1002	0.9	-0.3	0.9	-0.4	43.3	.62
1078	0.9	-0.2	0.8	-0.8	41.7	.63
1004	0.9	-0.3	0.7	-0.8	35.0	.64
1084	0.9	-0.3	0.7	-0.9	38.3	.64
1079	0.9	-0.4	0.7	-0.7	60.0	.60
1056	0.9	-0.5	0.8	-0.6	53.3	.62
1081	0.9	-0.7	0.7	-0.9	33.3	.66
1082	0.9	-0.8	0.8	-0.7	41.7	.66
1049	0.9	-0.7	0.7	-0.7	31.7	.66
1066	0.8	-0.8	0.8	-0.3	25.0	.64
1070	0.8	-1.2	0.6	-1.3	50.0	.67
1051	0.8	-0.6	0.3	-0.6	10.0	.54
1013	0.9	-1.2	0.6	-1.2	33.3	.70
1008	0.8	-1.0	0.5	-0.8	21.7	.66
1076	0.7	-1.3	0.5	-0.8	20.0	.67
1029	0.6	-1.8	0.4	-1.5	25.0	.73
1050	0.6	-2.7	0.5	-2.2	46.7	.76
1085	0.6	-2.1	0.4	-1.3	23.3	.74

To attempt to answer the third research question, dimensionality statistics were assessed. Table 5 is the dimensionality statistics. Figure 2 is a plot of the dimensionality provided by Winsteps. 44 percent of the variance is accounted for by the measures. Interestingly, there is a significant contrast that accounts for 5.71 percent of the variance. It is difficult to speculate on what this first contrast may represent. Perhaps, because paused transcription relies on writing for assessment, the contrast represents L2 writing competency. It is possible that students who are more competent in writing in English may perform better on this type of test. This is only speculative, and the contrast may be due to some other unidentified variable. Despite this significance of the contrast, this instrument appears to be sufficiently unidimensional. A look at the chart in Appendix E shows that there appears to be no systematic grouping of items that could account for additional variance and dependence.

Table 5
Dimensionality Statistics

	Eigenvalue	Observed	Expected
Total raw variance in observations	107.1	100.0%	100.0%
Raw Variance explained by measures	47.1	44.0%	44.0%
Raw variance explained by persons	16.9	15.8%	15.9%
Raw variance explained by items	30.1	28.2%	28.2%
Raw unexplained variance	60.0	56.0%	56.0%
Unexplained variance in 1st contrast	5.7	5.3%	

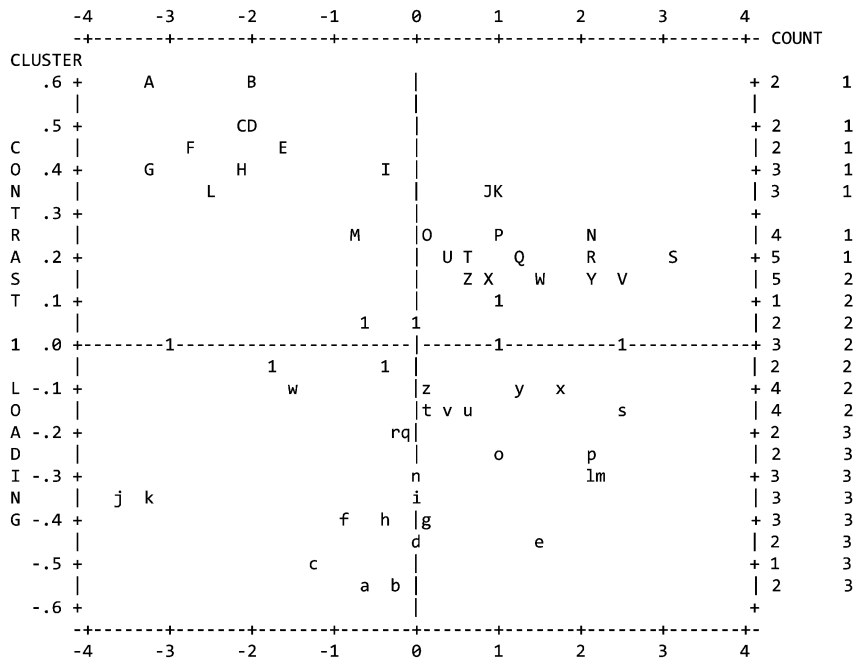

Figure 2. Dimensionality plot

The fourth research question was concerned with item dependence due to multiple instances of identical words found in this instrument. Table 6 shows items that highly correlated as an indicator of item dependence. The correlations in Table 4 show that multiple instances of identical items were not

dependent. However, there is considerable phrasal dependency. The successful transcription of a word seems to be most influenced by the words that precede it. For instance, Table 4 shows that if a participant were to not transcribe the word *To*, they almost certainly would not transcribe *Do.* This high phrasal dependence would seem problematic, but this test is meant to be a measure of listening ability. Phrasal dependence is likely a normal part of L2 listening. If an L2 listener does not hear a word, their probability of hearing a word that follows is severely limited. As such, controlling for this type of dependency is not practical or even desirable. Doing so would make this test a less efficient measure of L2 listening comprehension.

Table 6
Item Dependency Correlations

Dependent Items		*r*
F_To	C_Do	1.00
F_It	F_through	1.00
C_day	C_Is2	.75
F_Can	C_Help	.75
C_Very	C_Good	.74
F_My	C_Desk	.73
F_Where	F_If	.71
C_Mother	F_And	.70
F_The1	C_Subject	.69
C_Help	F_You2	.69
F_Can	F_You2	.68
F_I	F_Can	.68
F_And	C_Father	.66
F_What	C_Is1	.66
F_Their	F_Where	.65
C_Next	C_Week	.65
F_Before	C_Go	.65
F_The3	C_Live	.65
C_National	C_Holiday	.64

Limitations & Future Research

It should be remembered that this study is only meant as a field test and pilot study for an additional follow-on study that will utilize this instrument in a mixed effects method analysis. The sample size of this study was relatively small. All findings should be thought of as preliminary. While this study is a promising start to a validity argument for paused transcription, further research should be conducted to strengthen the argument. One approach that should be taken is to see how high the correlation is between a paused transcription test and a standardized norm referenced test such as the TOEFL. A high correlation between results on a paused transcription test and results of a listening section on a norm referenced test would strengthen the argument that the method is actually a measure of L2 listening proficiency and not a measure of some other phenomenon. Future research should be conducted to test the effect content has on student performance. The content of the test in this study was a teacher speaking about an upcoming assignment. It is possible that a test with content related to a context that is less familiar to students would affect comprehension and transcription rates.

Conclusion

The purpose of this study was to begin constructing a validity argument for paused transcription L2 listening tests. Specifically, this study was meant to field test and argue for the validity of an instrument that will be used in a future mixed effects model analysis to test the effects of various word attributes on successful L2 listening processing rates. The results of the Rasch analysis show that this method and specifically this iteration of the test meets the assumptions of Rasch analysis. The results also align with prior research by Field (2008) and Yeldham (2016) that showed content words are significantly favored over functors in L2 listening comprehension.

Paused transcription is a promising method that could give SLA researchers new insight into the nature of L2 listening processing. In the past, SLA researchers were limited to comprehension and recall tests to research how L2 listeners process incoming speech. Small research budgets and a lack of access to brain imaging technology have limited the methods that in-class L2 researchers can use. Paused transcription and other similar methods will open up new avenues of inquiry that will expand the field's understanding of what is occurring in the mind of a L2 listener.

References

Audacity - Free, open source, cross-platform audio software. (n.d.). Retrieved from https://www.audacityteam.org/

Bond, T., & Fox, C. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*(3rd ed.). New York, NY: Routledge.

Brown, C., Hagoort, P., & Ter Keurs, M. (1999). Electrophysiological signatures of visual lexical processing: Open- and closed-class words. *Journal of Cognitive Neuroscience,11*(3), 261-281. doi:10.1162/089892999563382

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*(1st ed., pp. 159-181). New York, NY: Routledge.

Ellis, R. (2008). *The study of second language acquisition* (2nd ed.). Oxford: Oxford University Press.

Field, J. (2008). Bricks or mortar: Which parts of the input does a second language listener rely on? *TESOL Quarterly,*411-432. doi:10.1002/j.1545-7249.2008.tb00139.x

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*(1st ed.). New York, NY: Routledge.

Graham, S., & Santos, D. (2013). Selective listening in L2 learners of French. *Language Awareness, 22*(1), 56-75. doi:10.1080/09658416.2011.652634

Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223.

Hauk, O., & Pulvermuller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology, 115*, 1090-1103. doi:10.1016/j.clinph.2003.12.020

Kuchinke, L., Vo, M., Hofmann, M., & Jacobs, A. (2007). Pupillary responses during lexical decision vary with word frequency but not emotional valence. *International Journal of Psychophysiology, 65*, 132-140. doi:10.1016/j.ijpsycho.2007.04.004

Linacre, J. M. (2019). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

MRC Psycholinguistic Database. (n.d.). Retrieved from http://websites.psychology.uwa.edu.au/school/MRCDatabase/uwa_mrc.htm

Rost, M. (2016). *Teaching and researching listening* (3rd ed.). New York, NY: Routledge. doi:10.1111/ijal.12003

Wright, B., & Linacre, J. (1994). Reasonable mean-square fit values. In *Rasch Measurement Transactions*(Vol. 8, p. 370). MESA Press.

Yeldham, M. (2016). The decoding of word classes by L2 English listeners. *英語教學,40*(1), 49-78. doi:10.6330/ETL.2016.40.1.03

Appendix A

Test Specification Table

Skill Focus	L2 Listening proficiency
Task Description	A short monologue in English will play. A tone followed by a 15 second pause will occur intermittently 12 times throughout the recording. When the participant hears the first tone, they will attempt to transcribe the last five words they heard preceding the tone. A second tone plays to inform participants that the monologue will begin again. Participants will attempt to transcribe 12 target phrases.
Task Purpose	The purpose of this test will be to examine the effects characteristics of words have on their successful processing. This is an extension of research by Field (2008) and Yeldham (2016) that demonstrated functors are not successfully processed by L2 listeners at the same rate as content words. The characteristics that will be tested for are lexical and prosodic stress, word length, frequency, part of speech, and word imageability. The results of the test will fill be examined using Rasch analysis to identify any items that are exhibiting odd behavior. In the subsequent study, the assessment will be analyzed using a mixed effects model.
Monologue Characteristics	All monologue will be grammatical. The language of the test should attempt to mimic naturalistic spoken speech. In order to ensure the monologue is schematically neutral, the content of the monologue will be a university English teacher speaking about upcoming class assignments.
Time	Approximately 10 minutes.
Materials	The audio file will be created using the software Audacity®. The audio file will begin with a speaker of the participant's native language reading the test instructions. This will be followed by the monologue. The monologue will be spoken by a native speaker of the participant's second language. The audio file will be embedded in a PowerPoint presentation that has the instructions for the test written in the participant's native language. Each student will be provided with a test form that has the instructions for the test written at the top. There will be twelve blanks on the form provided for transcription. An additional audio file with a practice phrase using similar language as the test will be created to familiarize participants with the test procedures.
Scoring Parameters	Dichotomously scored (successful or unsuccessful transcription). Misspelled but phonetically similar variants are counted as successful transcriptions. Each word is treated as a separate item. Each word is given a 0 for unsuccessful transcription and a 1 for successful transcription.
Instructions to Participants (English & Japanese)	You will hear a short monologue in English. This monologue is an English teacher talking about upcoming assignments. Within this monologue there are 12 beeps followed by pauses. When you hear this first beep, attempt to write the last five words you heard. You will then hear a second beep. This second beep means the monologue will begin again shortly. Write each phrase in English in the blank space provided on your test sheet. If you do not know the spelling of a word, try to write how the word sounds. Try to write exactly what you hear. You will have 15 seconds to write each phrase. これから短い英語の会話を聴いてもらいます。この会話の中では、英語の先生が宿題について話しています。会話の中では、１２回ブザー音が鳴ります。

	ブザー音の後には、5 つの単語が聴こえます。テスト用紙の空欄に、ブザー音の後に聴こえた５つの英単語を記入してください。それぞれのブザー音の後に、再度ブザー音が鳴りますが、これは次の会話が始まりまる合図のブザー音です。スペルがわからない場合も、空欄にするのではなく、聴こえた音に合わせてスペルを綴るようにしてください。記入する時間はそれぞれ１５秒ずつあります。できる限り、聴き取った通りの単語を記入するようにしてください。
Item Example	The underlined excerpt from an audio recording is the target phrase. – “Soon you will have to submit an outline of your essay.” (Tone)(15 second pause)
Test Procedure	Participants will be informed they are taking a listening test for research purposes. Before the test form is administered, the practice phrase audio file will be played. The test administrator will model transcribing the test phrase on the board. They will also model phonetically transcribing the word if spelling is unknown. Next, the test form will be administered. When all participants have received the test form, the test audio file will be played. The test administrator will standby during the test to ensure participants remain quiet.

Appendix B

Test Monologue

Underlined Sections are the target phrases

Alright everyone, please listen up. Before you leave, we will discuss your assignments. You have a lot (1) of work to do soon. You have many deadlines that you need to remember. The most important thing is the essay. (2) Next week send it through email to me. This isn’t your first paper, so it should be easy for you. (3) The subject of the essay is discussing your family and home. Be sure to tell me (4) about your mother and father. What are their jobs? (5) What is their personality like? You can also talk about your brothers and sisters. Use lots of details. You should also discuss (6) the city where you live. The essay should be 1000 words and is due by Friday. On Wednesday, we (7) will have the grammar test. The test will be on the grammar we studied in chapter five. After you finish the test, (8) I can help you with editing your essay. Just bring a copy and I’ll work with you. On Thursday, there is no class because that (9) day is a national holiday. You can still reach me through email, (10) if there is a question about the essay you need answered. I finished grading your quizzes. Grab them from (11) my desk before you go. This quiz wasn’t so difficult, so most of you (12) guys got very good grades. I hope you have a good weekend.

Appendix C

Test Form

Name _____ Number _____

You will hear a short monologue in English. This monologue is an English teacher talking about upcoming assignments. Within this monologue there are 12 beeps followed by pauses. When you hear this first beep, attempt to write the last five words you heard. You will then hear a second beep after 15 seconds. This second beep means the monologue will begin again shortly. Write each phrase in English in the blank space provided on this sheet. If you do not know the spelling of a word, try to write how the word sounds. Try to write exactly what you hear. You will have 15 seconds to write each phrase.

1. _____
2. _____
3. _____
4. _____
5. _____
6. _____
7. _____
8. _____
9. _____
10. _____
11. _____
12. _____

Appendix D

Test Administration Instructions

1. Tell the students they will be doing a quick listening quiz for research purposes. Explain they will be listening to a monologue of a teacher talking about essays and homework. Whenever they hear a beep, they must try to write the last five words they heard.
2. Before handing out the sheets, tell them you will give an example.
3. Go through the practice slide in the PowerPoint. The slide has an audio file and explains that students must write the last 5 words they hear after a beep and that misspelling are ok. Explain they will have 15 second to write each phrase.
4. Give the students the test sheet after you finish the practice slide.
5. Tell the students to write their name and number. Also, tell them to remain quiet during the test.
6. Begin the audio file on the next PowerPoint slide. The instructions for the test will play at the beginning of the file. A few seconds after the instructions, the monologue will begin. Wait for the test to finish. Ensure students remain quiet and on task.
7. Collect the test sheet once the test is complete. Ensure the students remembered to write their name and number.