SHIKEN

Volume 23 • Number 2 • December 2019

Contents

Testing and Evaluation SIG Newsletter

ISSN 1881-5537

Shiken

Volume 23 No. 2
December 2019

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.

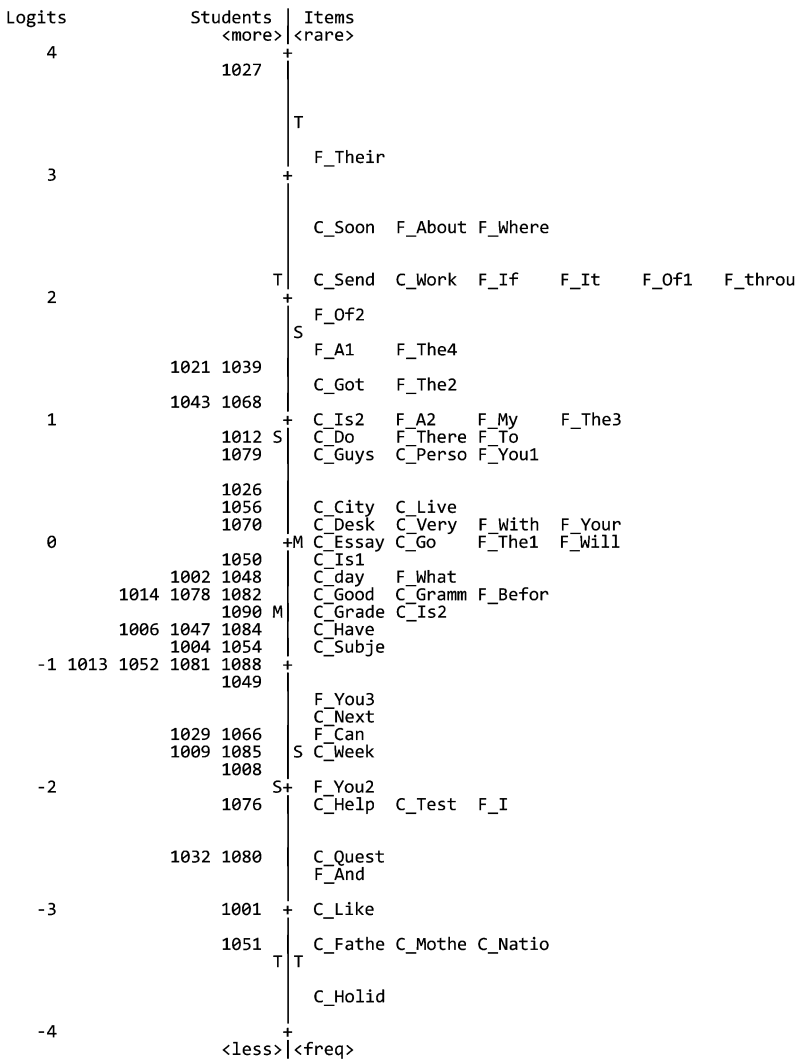

Figure 1. Wright map of relationship between item difficulty and participant ability.

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%	

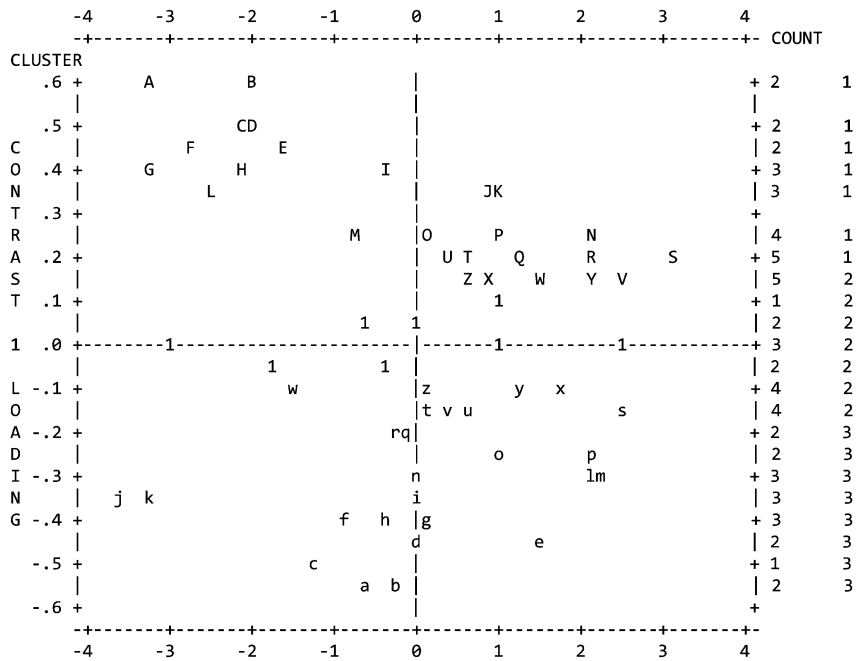

Figure 2. Dimensionality plot

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.

An Analysis of Vocabulary Level in Reading Passages of the National Center Test

Ewen MacDonald
macdonald-e@kanda.kuis.ac.jp
Kanda University of International Studies

Abstract

A strong knowledge of high frequency vocabulary is essential for second language (L2) learners of English as it provides a large proportion of text coverage for general English texts. Additionally, a minimum understanding of 95% to 98% of the words in a text is essential to allow for reasonable comprehension and guessing of words in context. The current study examines the text coverage provided by the New General Service List (NGSL), a list of high frequency vocabulary for L2 English learners, for reading comprehension passages of Japan's National Center Test, a standardised high-stakes national university entrance examination. Results showed that text coverage provided by the NGSL exceeded the minimum 95% threshold for reasonable reading comprehension for examinations between 2015 and 2019 with a similar level of vocabulary found across these years. The findings present a compelling argument that supporting students to acquire high frequency vocabulary should be strongly focused on at secondary school in Japan.

Keywords: Center Test, New General Service List, high frequency vocabulary, vocabulary level, Japan entrance examinations

A strong knowledge of high frequency vocabulary is foundational to second language (L2) learning and obtaining general language proficiency, as words that occur with a high level of frequency provide a disproportionately high percentage of text coverage (Stoeckel & Bennett, 2015). An analysis by Nation (2006) of the British National Corpus showed that knowledge of 3,000 to 4,000 word families is necessary to understand 95% of a general English text. This is an important minimum vocabulary threshold, with Laufer (1989) estimating that knowing 95% of the words in a text allows reasonable comprehension and guessing the meaning of unknown words in context. Later studies concluded that while 95% coverage allows minimally acceptable comprehension, text coverage of at least 98% based on knowledge of 8,000 word families is ideal, and should be an eventual goal for learners as it allows accurate guessing of unknown words in context and unassisted reading for pleasure (Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang & Grabe, 2011). It can therefore be said that a sound knowledge of core vocabulary is essential for L2 learners of English, and the acquisition of high frequency vocabulary should be focused on in teaching and learning English as a second language.

The New General Service List (NGSL)

With the importance of this in mind, Browne, Culligan, and Phillips (2013) published the New General Service List (NGSL) which contains 2,801 important high frequency words useful for second language learners of English. The list was created after an analysis of a 273 million-word subsection of the Cambridge English Corpus and provides 90.34% coverage of this corpus. It has the aim of providing the highest possible coverage of English texts with the fewest words possible. A "modified lexeme" approach is used to group words in the list with a headword including all its parts of speech and inflected forms, but not including derived forms with non-inflection suffixes (Browne, 2014). This results in the NGSL containing 2,368 word families and 2,801 modified lemmas.

Vocabulary in Japanese University Entrance Examinations and Textbooks

The majority of students in Japan take university entrance examinations in their final year of high school, with their scores having a major bearing on receiving admission offers to universities. English examinations include a section which tests reading comprehension and knowledge of vocabulary and

grammar. These examinations have faced criticism due to their difficulty level, and the large vocabulary size required that is greater than what is expected of high school students (Chujo & Hasegawa, 2004; Hasegawa, Chujo, & Nishigaki, 2004; Kikuchi, 2006). This has had a strong "washback effect" on English education and vocabulary learning.

Although English reading materials in Japanese high school textbooks are often considered difficult for students (Browne, 1998), vocabulary in these textbooks is still often insufficient in preparing students for the demands of university entrance examinations. The lexical coverage of vocabulary in textbooks for many examinations has been found to be low (Chujo, 2004; Chujo & Hasegawa, 2004; Hasegawa, Chujo, & Nishigaki, 2006; Kitao & Kitao, 2011; Matsuo, 2000; Underwood, 2010). This can lead to test-takers facing great difficulty in adequately comprehending the reading passages of examinations as they regularly encounter unknown words while reading beyond their level, potentially leading to demotivation and a lack of confidence. Arguments have been made to bring the vocabulary used in examinations in line with what students actually study in school (Hasegawa, Chujo, & Nishigaki, 2006; Matsuo, 2000).

With the vocabulary in textbooks often inadequate for meeting the demands of entrance examinations, Japanese high school students are compelled to spend a large amount of time studying and memorising hundreds of additional low frequency words for intensive reading purposes. This is done by using a corpus developed from past entrance examinations and published through cram schools, which covers a large amount of low frequency vocabulary (Underwood, 2010). Hence, the vocabulary load on students is very heavy, and a concerted effort is required to memorise low frequency words by students who wish to gain entrance to prestigious universities.

This focus on learning low frequency vocabulary can lead to a lack of high frequency vocabulary knowledge. While the average vocabulary size of Japanese college students was previously estimated at 3,715 word families, students appeared to have "a consistent lack of knowledge of even the most frequent words of English" (McLean, Hogg & Kramer, 2014, p. 53) which can make comprehension more difficult. Ideally, if high school students were required to learn the most frequent vocabulary, it would allow them to comprehend a higher proportion of English texts. This is important as there is a strong relationship between L2 vocabulary knowledge and reading comprehension, with 95% understanding of the words in a text necessary for minimally acceptable comprehension, and 98% for most learners to read unassisted (Hu & Nation, 2000). Laufer and Ravenhorst-Kalovski (2010) note that even small increments in vocabulary knowledge can contribute to reading comprehension. If students were able to recognise high frequency words immediately, it would also lead to faster and more fluent reading (Underwood, 2010).

Vocabulary Requirements for the National Center Test

There are two types of university entrance examinations in Japan; university-specific examinations offered by public and private universities, and the National Center Test (NCT) for University Admissions, a standardised national examination administered once a year and taken by over 500,000 students (National Center for University Entrance Examinations, 2019a). A variety of subjects are tested with English as one of the required subjects. The English examination includes both listening and reading, with the reading section containing several reading comprehension passages with multiple choice questions.

The English examination of the NCT has a reputation for being easier than those of public and private universities. Several studies found the readability and vocabulary level of the NCT to be appropriate for Japanese high school students at their time of graduation, while the difficulty of individual university entrance examinations were significantly above students' expected level (Chujo, 2004; Chujo, & Hasegawa, 2004; Hasegawa, Chujo, & Nishigaki, 2006; Makoto, MacGregor, Nakajima, & Omori, 2006; Matsuo, 2000). In addition, the vocabulary in junior and senior high school (JSH) textbooks was found to provide coverage of at least 95% for vocabulary in several NCTs (Chujo, & Hasegawa, 2004; Hasegawa,

Chujo, & Nishigaki, 2006). This suggests that the NCT is a more suitable examination for Japanese high school graduates.

More recent research similarly found the NCT to be less challenging than individual university examinations, but also suggested the readability and lexical difficulty level of the test had increased. Tani (2008) reported that 91.2% of the vocabulary in the 2008 NCT was covered in high school textbooks, while Kitao and Kitao (2011) found that textbooks provided only 81.26% lexical coverage (2,368 words) for the 2010 NCT. Although still easier than other entrance examinations, they mentioned that the difficulty level of reading passages had increased since the 2006 NCT, with more low frequency vocabulary and a higher readability level. This was further supported by Underwood (2010) who noted that the readability of the 2008-2009 NCTs had increased to a level comparable to entrance examinations of prestigious public and private universities. These findings suggest that although the NCT is still the most appropriate examination, there is an increased likelihood that many test-takers may have significant difficulty comprehending the reading passages of the NCT.

A strong overlap has been observed when comparing high frequency word lists with vocabulary contained in JSH textbooks and the NCT. Using a lemmatised high frequency word list made from the British National Corpus (BNC), Chujo (2006) discovered the top 3,100 words from the BNC covered 95% of vocabulary in the 2001-2002 NCTs as well as vocabulary in JSH textbooks. A later study by Underwood (2010) found the 2,000 words of the General Service List (GSL), the predecessor of the NGSL, provided 92.67% vocabulary coverage of the 2003-2007 NCTs, 86.45% of the 2008-2009 NCTs and 88.59% for final grade high school textbooks, indicating a degree of compatibility between them. When adding an additional 570 word families from the Academic Word List (AWL) to the GSL, vocabulary coverage for the NCTs increased to over 93%. Underwood noted that this observation was a promising trend in mitigating the difficulty of the NCT and helping students retain high frequency vocabulary knowledge; yet, it would be contingent on choosing this vocabulary for learning at high school level in Japan.

The importance for L2 learners to acquire knowledge of high frequency words and its positive effect on reading comprehension is clearly understood. However, with gaps in their core vocabulary knowledge, Japanese students often have difficulty when sitting university entrance examinations, such as the NCT. With a lack of recent studies and little research comparing high frequency word lists with the NCT, this study uses the NGSL to analyse the vocabulary level of past NCTs, and sets out to answer the following questions:

1. What percentage of vocabulary coverage does the NGSL provide for long reading comprehension passages of the 2014-2019 NCTs?
2. Does this vocabulary coverage achieve the ideal 98% text coverage to allow comfortable reading comprehension and guessing of words in context?

Method

Data Collection

Long reading comprehension passages (Sections 4, 5 and 6) from the 2014-2019 NCTs were analysed using an online corpus analysis tool called VocabProfile (Cobb, n.d.). Using this tool, a text can be entered or uploaded, and the number of words that the text contains from frequency bands of different word lists is calculated, with statistical data being generated. In this study, the reading passages of past tests were copied and pasted into VocabProfile.

Analysis

The analysed passages were compared against the first three NGSL bands, with the text coverage provided by the NGSL calculated for each frequency band.

- NGSL 1 (first 1000 modified lemmas)
- NGSL 2 (second 1000 modified lemmas)
- NGSL 3 (third 801 modified lemmas)
- Off-List words (words not contained in the NGSL)

For the purpose of this study, with the assumption that reading comprehension would not be negatively affected, the following types of words that students would be expected to easily understand from context were recategorised as NGSL Band 1 words: Japanese words, abbreviations, short exclamations, proper nouns, ordinal numbers, days of the week and months.

Results

Table 1 shows the frequency level of the first three bands of the NGSL and off-list words, the percentage of vocabulary from the reading passages covered by each band in the NCTs from 2014 to 2019, and the cumulative text coverage percentage.

Table 1
NGSL text coverage for reading passages of the 2014-2019 NCTs

	2014 NCT		2015 NCT		2016 NCT		2017 NCT		2018 NCT		2019 NCT	
Frequency Band	TC %	Cu %	TC %	Cu %	TC %	Cu %	TC %	Cu %	TC %	Cu %	TC %	Cu %
NGSL 1	85.64	85.64	88.49	88.49	81.60	81.60	85.91	85.91	81.44	81.44	85.17	85.17
NGSL 2	6.60	92.24	6.73	95.22	9.59	91.19	7.61	93.52	9.68	91.12	8.09	93.26
NGSL 3	2.38	94.62	1.71	96.93	5.30	96.49	2.91	96.43	4.93	96.05	2.51	95.77
Off-List	5.38	100.00	3.07	100.00	3.51	100.00	3.57	100.00	3.95	100.00	4.23	100.00

Note. NCT = National Center Test, NGSL = New General Service List, Off-List = Words not contained in the NGSL, TC % = Text coverage percentage provided by the NGSL, Cu % = Cumulative text coverage percentage provided by the NGSL

It can be seen that the 2,801 high frequency words from the NGSL provided between 95% to 97% overall text coverage of the NCTs from 2015 to 2019. This is above the minimum 95% vocabulary threshold to allow for reasonable reading comprehension and guessing of words in context, but not quite the 98% required for comfortable comprehension. In the 2014 examination, the NGSL provided overall coverage marginally below the 95% threshold, indicating that more low frequency vocabulary was used in this year.

Examinations from 2015 to 2019 also contained a very similar level of vocabulary, with only a 0.88% difference in the cumulative percentage at the NGSL 3 frequency band. For the 2015 NCT, knowledge of only the top 2,000 words of the NGSL was sufficient to reach the 95% threshold. However, this was not repeated in examinations from other years.

A further breakdown of the examinations shows variability in the level of vocabulary between reading passages (see Appendix A). For example, in the 2018 NCT, the NGSL provided over 98% text coverage of the reading passage in Section 4, but less than 95% coverage of the reading passage in Section 6.

Discussion

When comparing the findings with the results of past studies, an observation can be made that the amount of high frequency vocabulary contained in the reading passages of the NCT has significantly increased in recent examinations, continuing the trend identified by Underwood (2010).

The 2,801 word NGSL has provided 96% to 97% coverage of the NCT since 2015, above the 95% threshold for reasonable comprehension and close to the 98% threshold for more comfortable comprehension and accurate guessing of words in context. Therefore, a strong case can be made that students would greatly benefit by learning the high frequency vocabulary in the NGSL. The coverage provided for individual reading passages in different examinations varied from 93% to 98%, showing that knowledge of high frequency vocabulary would help students comfortably comprehend some texts, but they may potentially still struggle with other passages.

According to the New Course of Study guidelines for foreign language education in Japan set by the Ministry of Education, Sports, Culture, Science and Technology (MEXT), students are expected to learn 3,000 words by the end of high school (MEXT, 2015). Hence, it can be argued that learning the words in the NGSL could be considered a realistic goal with consistent and targeted vocabulary instruction and learning. The acquisition of high frequency words and helping students develop a strong retention of them should, therefore, be focused on more in secondary education through regular exposure and practice, and by including a greater amount of this vocabulary in JSH textbooks.

Having knowledge of high frequency words would help students more easily comprehend English texts, such as those in the NCT, and increase their reading fluency in turn. Regardless of whether the NGSL or any other high frequency word list is used, there is a clear need for high school students to acquire core vocabulary, not only for improving their reading comprehension for entrance examinations, but also for their general English language proficiency.

As university entrance examinations of public and private universities have previously been found to contain a higher level of vocabulary than that of the NCT, high school students will undoubtedly be required to continue learning low frequency vocabulary independently through lexical corpora in cram schools. For the learning of high frequency vocabulary to be implemented most effectively, the vocabulary in other university entrance examinations would need to be more closely aligned with that in the NCT and in JSH textbooks. Learning both high and low frequency words for both the NCT and other individual university examinations would increase the already excessive amount of time students need to spend on studying vocabulary. If students cannot cope with such a heavy vocabulary load, this could potentially result in knowledge gaps of high frequency vocabulary remaining.

Conclusion

This study examined the vocabulary level of reading comprehension passages of the NCT by calculating the text coverage provided by the NGSL. The NGSL was found to give greater than 95% text coverage for passages in the tests from 2015 to 2019, indicating that receptive knowledge of high frequency vocabulary would allow students to reasonably comprehend the reading passages of the NCT and increase the likelihood of correctly guessing unknown words in context. The level of high frequency vocabulary was also found to be consistent across the tests between 2015 and 2019 and significantly higher than in previous years.

An important limitation of this study is that vocabulary knowledge is not the only factor that can affect reading comprehension. Other factors should be considered when determining the difficulty level for students to comprehend reading passages; e.g. readability measures, the length and complexity of words,

sentences and texts, students' grammatical knowledge, background knowledge of the topic, level of confidence and automaticity, as well as students' knowledge of reading strategies.

Additionally, it should be noted that proper nouns, often presumed to be easily understood by L2 learners, were recategorised in VocabProfile as NGSL Band 1 words when analysing the reading passages. Brown (2010) points out that studies of text coverage have differed in their treatment of proper nouns with regard to their inclusion or exclusion from data analysis which can lead to different findings. For the purpose of this study, it was assumed that at the time of undertaking the NCT, 3rd year Japanese high school students ought to be able to recognise proper nouns and hence they would not be problematic.

Further research should be undertaken on the text coverage provided by high frequency vocabulary lists for JSH textbooks and for different public and private university entrance examinations. A study by Kaneko (2013) found that the 2,570 word items provided by the GSL and the AWL provided 95.29% average text coverage of the reading passages in the 2003-2011 entrance examinations of Tokyo University, a highly prestigious school. This suggests that even public and private entrance examinations may now contain a larger amount of high frequency vocabulary than previous studies have shown, and that learning high frequency words may also be beneficial for such examinations.

In addition, ongoing changes are being proposed and made by MEXT to the New Course of Study guidelines and the university entrance examination system, including whether to utilise English tests offered by private sector organisations (National Center for University Entrance Examinations, 2019b, 2019c). The effects of any future changes on the vocabulary found in examinations and school textbooks, the teaching of vocabulary in the classroom, and the overlap with vocabulary contained in high frequency vocabulary lists should continue to be examined.

References

Brown, D. (2010). An improper assumption? The treatment of proper nouns in text coverage counts. *Reading in a Foreign Language, 22*(2), 355-361. Retrieved from https://files.eric.ed.gov/fulltext/EJ901552.pdf

Browne, C. (1998). Japanese high school textbooks: Help or hindrance? *Temple University Japan Working Papers in Applied Linguistics, 12*, 1-13. Retrieved from http://www.charlie-browne.com/wp-content/downloadable-files/JapaneseHighSchoolTexts.pdf

Browne, C. (2014). A New General Service List: The better mousetrap we've been looking for? *Vocabulary Learning and Instruction, 3*(1), 1-10. doi:10.7820/vli.v03.1.browne

Browne, C., Culligan, B., & Phillips, J. (2013). The New General Service List. Retrieved from http://www.newgeneralservicelist.org

Chujo, K. (2004). Measuring vocabulary levels of English textbooks and tests using a BNC lemmatised high frequency word list. In J. Nakamura, N. Inoue, & T. Tabata (Eds.), *English corpora under Japanese eyes* (pp. 231-249). Amsterdam, Netherlands: Rodopi.

Chujo, K., & Hasegawa, S. (2004). Goi no cover ritsu to readability kara mita daigaku eigo nyushi mondai no nanido [Assessing Japanese college qualification tests using JSH text coverage and readability indices]. *Nihon University Student Faculty of Engineering Research Report B, 3*7, 45-55. Retrieved from http://www.cit.nihon-u.ac.jp/laboratorydata/kenkyu/publication/journal_b/b37.5.pdf

Cobb, T. (n.d.) *VocabProfile* [Computer program]. Retrieved from http://www.lextutor.ca/vp/

Hasegawa, S., Chujo, K., & Nishigaki, C. (2006). Daigaku nyushi eigo mondai goi no nanido to yuuyousei no jidaiteki henka [A chronological study of the level of difficulty and the usability of the

English vocabulary used in university entrance examinations]. *JALT Journal, 28*(2), 115-134. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/jj-28.2-art1.pdf

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430. Retrieved from https://nflrc.hawaii.edu/rfl/PastIssues/rfl131hsuehchao.pdf

Kaneko, M. (2013). Estimating the reading vocabulary-size goal required for the Tokyo University entrance examination. *The Language Teacher, 37*(4), 40-45. Retrieved from https://jalt-publications.org/files/pdf-article/37.4tlt_art1.pdf

Kikuchi, K. (2006). Revisiting English entrance examinations at Japanese universities after a decade. *JALT Journal, 28*(1), 77-96. Retrieved from https://jalt-publications.org/sites/default/files/pdf-article/jj-28.1-art5.pdf

Kitao, K., & Kitao, S. K. (2011). Readability and vocabulary level of reading passages in Japanese university entrance exams. *Journal of Culture and Information Science, 6*(1), 11-20. Retrieved from https://doors.doshisha.ac.jp/duar/repository/ir/15832/039000060002.pdf

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. H. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Clevedon, UK: Multilingual Matters.

Laufer, B., & Ravenhorst-Kalvoski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30. Retrieved from https://nflrc.hawaii.edu/rfl/April2010/articles/laufer.pdf

Makoto, H., MacGregor, L., Nakajima, K., & Omori, Y. (2006). Daigaku eigo nyushi mondai no chousa bunseki [Reading passages in entrance examinations to Japanese universities]. *Language, Culture and Society, 6*, 139-184. Retrieved from https://glim-re.repo.nii.ac.jp/?action=repository_uri&item_id=1108&file_id=22&file_no=1

Matsuo, H. (2000). An analysis of Japanese high school English textbooks and university entrance examinations: A comparison of vocabulary. *Annual Review of English Language Education in Japan, 11*, 141-150. doi:10.20581/arele.11.0_141

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction, 3*(2), 47-55. doi:10.7820/vli.v03.2.mclean.et.al

MEXT (2015). Gaikokugoka gaikokugo katsudo ni okeru mokuhyo shido naiyo to [Objectives and guidance for foreign language courses and activities]. Retrieved from http://www.mext.go.jp/b_menu/shingi/chukyo/chukyo3/056/siryo/__icsFiles/afieldfile/2015/10/29/1363262_10.pdf

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59-82. doi:10.1353/cml.2006.0049

National Center for University Entrance Examinations (2019a). Center shiken shiganshasu jukenshashu heikin ten no suii [Number of examinees and average scores (main exam) since 2018 Center Test]. Retrieved from https://www.dnc.ac.jp/center/suii/index.html

National Center for University Entrance Examinations (2019b). Daigaku nyugaku kyotsu test (shin test) tou nitsuite [About Common Test for University Admissions (new test)]. Retrieved from https://www.dnc.ac.jp/daigakunyugakukibousyagakuryokuhyoka_test/

National Center for University Entrance Examinations (2019c). Daigaku nyushi eigo seiseki teikyo shisutemu [University entrance examination English provision system]. Retrieved from https://www.dnc.ac.jp/eigo_seiseki_system/

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43. doi:10.1111/j.1540-4781.2011.01146.x

Stoeckel, T., & Bennett, P. (2015). A test of the New General Service List. *Vocabulary Learning and Instruction, 4*(1), 1-8. doi:10.7820/vli.v04.1.2187-2759

Tani, K. (2008). Daigaku nyushi Center shiken goi to koukou eigo kyoukasho no goi hikaku bunseki [A comparative analysis of the National Center Test vocabulary and high school English textbook vocabulary]. *Practical English Studies, 14*, 47-55. doi:10.11200/japeronso1991.2008.47

Underwood, P. (2010). A comparative analysis of MEXT English reading textbooks and Japan's National Center Test. *RELC Journal, 41*(2), 165-182. doi:10.1177/0033688210373128

Appendix

Table A1
NGSL text coverage for individual reading passages of the 2019 NCT

Frequency	Section 4		Section 5		Section 6	
Band	TC %	Cu %	TC %	Cu %	TC %	Cu %
NGSL 1	84.06	84.06	88.99	88.99	81.21	81.21
NGSL 2	10.80	94.86	3.32	92.31	11.90	93.11
NGSL 3	2.06	96.92	3.02	95.33	2.24	95.35
Off-List	3.08	100.00	4.67	100.00	4.65	100.00

Note. NGSL = New General Service List, Off-List = Words not contained in the NGSL, TC % = Text coverage percentage provided by the NGSL, Cu % = Cumulative text coverage percentage provided by the NGSL

Table A2
NGSL text coverage for individual reading passages of the 2018 NCT

Frequency	Section 4		Section 5		Section 6	
Band	TC %	Cu %	TC %	Cu %	TC %	Cu %
NGSL 1	83.42	83.42	79.47	79.47	82.28	82.28
NGSL 2	12.89	96.31	9.09	88.56	8.28	90.56
NGSL 3	2.11	98.42	8.31	96.87	3.15	93.71
Off-List	1.58	100.00	3.13	100.00	6.29	100.00

Table A3
NGSL text coverage for individual reading passages of the 2017 NCT

Frequency	Section 4		Section 5		Section 6	
Band	TC %	Cu %	TC %	Cu %	TC %	Cu %
NGSL 1	81.18	81.18	85.28	85.28	90.53	90.53
NGSL 2	11.76	92.94	8.12	93.40	3.44	93.97
NGSL 3	2.35	95.29	2.20	95.60	4.13	98.10
Off-List	4.71	100.00	4.40	100.00	1.90	100.00

Table A4
NGSL text coverage for individual reading passages of the 2016 NCT

Frequency Band	Section 4		Section 5		Section 6	
	TC %	Cu %	TC %	Cu %	TC %	Cu %
NGSL 1	74.35	74.35	86.62	86.62	81.87	81.87
NGSL 2	13.36	87.71	8.09	94.71	8.22	90.09
NGSL 3	7.11	94.82	2.35	97.06	7.10	97.19
Off-List	5.18	100.00	2.94	100.00	2.81	100.00

Table A5
NGSL text coverage for individual reading passages of the 2015 NCT

Frequency Band	Section 4		Section 5		Section 6	
	TC %	Cu %	TC %	Cu %	TC %	Cu %
NGSL 1	88.46	88.46	93.97	93.97	83.09	83.09
NGSL 2	6.49	94.95	2.85	96.82	10.63	93.72
NGSL 3	1.44	96.39	1.17	97.99	2.42	96.14
Off-List	3.61	100.00	2.01	100.00	3.86	100.00

Table A6
NGSL text coverage for individual reading passages of the 2014 NCT

Frequency Band	Section 4		Section 5		Section 6	
	TC %	Cu %	TC %	Cu %	TC %	Cu %
NGSL 1	82.37	82.37	88.83	88.83	85.11	85.11
NGSL 2	9.74	92.11	4.85	93.68	6.43	91.54
NGSL 3	1.05	93.16	3.88	97.56	1.72	93.26
Off-List	6.84	100.00	2.44	100.00	6.74	100.00

Grit and Intrinsic Motivation for Language Learning: Instrument validation using the Rasch model

Michael J. Giordano
mikegio123@gmail.com
Kwansei Gakuin University

Abstract

University student grit and intrinsic motivation were measured by survey to determine the viability of grit as a language learning construct in this pilot study. The instruments were analyzed using the Rasch model in Winsteps (version 4.0) for construct validity, person and item fit, and unidimensionality. Logit scores were used in a correlation analysis to determine the relationship between grit, intrinsic motivation, and autonomous language learning dependent variables. The sample-dependent results suggest that the grit and intrinsic motivation instruments are unidimensional. Additionally, they are moderately correlated (r =.40) but not significantly related to autonomous language learning measurements. However, small sample size and poor person reliability limit the generalizability of the results. Person reliability and item targeting of the two constructs are discussed and disattenuated correlations are used to suggest avenues for future research.

Keywords: Rasch model, instrument validation, grit, motivation

Grit, as defined in the previous literature, is a psychological construct comprised of "a perseverance and passion for long-term goals" (Duckworth, Peterson, Matthews, & Kelly, 2007, p. 1087). Duckworth et al. (2007) originally hypothesized that the individuals who are grittier than others would view achievement as a marathon and would be able to power-through boredom and overcome hardship more than less gritty individuals. This construct, if applicable, would fit in well as an additional individual difference variable alongside motivation, willingness to communicate, self-efficacy, or anxiety in second language (L2) learning. That being said, little research has been conducted on the relationship between grit, motivation, and language learning.

Grit as a construct has been evaluated many times in educational, psychological, sociological research, though it has rarely been examined in foreign language research. In their original study, Duckworth et al. (2007) evaluated the construct validity using confirmatory factory analysis to create the original 12-item grit scale. They discovered a two-factor construct; consistency of interest and persistence of effort. These items were further refined to make a short grit scale (Grit-S) containing only eight items yet effectively and more efficiently measuring the same construct and facets (Duckworth & Quinn, 2009). Additionally, they discovered that grit related significantly to each of the Big Five personality traits; grit is highly positively related to conscientiousness, negatively related to neuroticism, and weakly correlated with agreeableness, extraversion, and openness to experiences. In summary of these two studies, grittier adults progress farther in their education and make fewer career changes; grittier adolescents earn higher GPAs; grittier West Point cadets are less likely to drop out after the first training session; and grittier spelling bee finalists are more likely to advance to later rounds (Duckworth & Quinn, 2009, p. 172-173). These original studies sparked a plethora of research into grit as predictor of success and achievement.

Duckworth and Eskreis-Winkler (2013) argue that one mechanism of grit is deliberate practice. Those people who are able to suffer through the challenges, focus intently on a long-term goal, and repeatedly put in the effort are the most successful. Deliberate practice has been argued by DeKeyser (2007) to be a fundamental aspect of automatization and overall second language acquisition. It seems natural then that grit should also be considered an L2 individual difference (ID) variable alongside other variables like anxiety, willingness to communicate, and, perhaps most similarly, motivation.

Learning a language is a lifelong pursuit with many ups and downs. Students and language learners of all ages are often bombarded with a variety of achievement goals and standards and may suffer fluctuations in motivation. Deci and Ryan (1985), in their discussion of Self-Determination Theory, suggest that motivation for human behavior and the regulation of that motivation is divided into categories and factors. Some people are motivated to act through external pressure or fear of being watched (extrinsic motivational factors). Whereas others are motivated by self-authored or internal factors (intrinsic motivational factors). Those people who are more intrinsically motivated toward a goal or an action tend to show "more interest, excitement and confidence, which in turn is manifest both as enhanced performance, persistence, and creativity and as heightened vitality, self-esteem and general well-being" than those who are extrinsically motivated (Ryan & Deci, 2000, p. 69). The connections between grit—perseverance and passion for long-term goals—and intrinsic motivation are plentiful. Self-driven people who accomplish long-term goals are likely to be both gritty and intrinsically motivated. The extent to which grit and intrinsic motivation are related is the intended outcome of this validation pilot study.

Method

Participants

Thirty-nine first-year students (F = 34, M = 5) from a private university in western Japan participated in this study by responding to a questionnaire via Google Forms. Of these 39 students, two needed to be removed from the study due to careless answering. This reduced to total convenience sample to 32 women and five men ($N = 37$) in three intact classes, including students who were no longer enrolled in the course. Their enrollment in the course did not affect their ability to respond or the presence of the two constructs. Their enrollment status only affected my ability to compare their grit and intrinsic measurements with their autonomous language learning scores on English Central. English Central is an online language learning tool designed to provide learners with videos, vocabulary learning management and quizzes, and speech and pronunciation practice ("English Central", n.d.). Participants in this study were required to watch five videos, study 50 new vocabulary items, and speak 50 lines every week using their personal computer and smartphone. The Google Form containing the questionnaire was available on the Learning Management System (LMS). Participants also received an email notification with the same message and reminder in class to complete the questionnaire in their free time. Ten minutes at the end of one class was allocated to help students access and complete the questionnaire if they wished.

Instrument

Grit items. The grit items on the instrument were developed by borrowing heavily from Duckworth and Quinn's (2009) Short grit (Grit-S) survey and—taking Kramer, McLean, and Martin's (2017) suggestions to heart—were altered to specifically represent language learning grit. All items on the Grit-S survey appear in a similarly worded version in the same order in the current questionnaire. Negatively keyed items which appear in the original survey also appear in the altered version for the current study, altered to better represent the construct of language learning grit rather than general grit. For instance, the first item on the original survey, *New ideas and projects sometimes distract me from previous ones.* was changed to *New assignments or projects in my language class distract me from older assignments or projects.* Two additional items designed to measure participant grittiness were added to the end of the questionnaire to help separate students into groups of high, medium, and low grit. These two additional items were: *When I am studying a language, I cannot be distracted from my task* and *I get disappointed and give up when I am unsuccessful.* In total, there were 10 language learning grit items including five reverse or negatively keyed items.

Intrinsic items. The inspiration for the intrinsic motivation items came from a study regarding intrinsic, extrinsic and integrative motivation by Noels, Clément, and Pelletier (2001). Many of the intrinsic motivation items in the current study were adapted and altered to make "I" statements of varying degrees of difficulty to represent varying degrees of intrinsic motivation. Some examples of the intrinsic motivation items include, *I enjoy learning new things in another language* and *I want to be someone who can speak more than one language.* None of the intrinsic motivation items were reversely keyed following the advice of previous research (Schmitt & Stults, 1985; Swain, Weathers, & Neidrich, 2008; Bond & Fox, 2015; Credé, 2018). Both constructs were measured using a five-point Likert-like scale, ranging from *Exactly like me* to *Not like me at all.* An odd numbered scale was chosen to match the original grit instrument. Construct maps and Item specifications can be found in Appendix A and B.

Dependent variables. In addition to the grit scale and intrinsic motivation measurement scores in the questionnaire, three additional measurements were included for each participant where available: English Central video count, English Central vocabulary count, and English Central spoken lines count. As part of the participants' weekly homework assignments, they are required to watch five videos, speak 50 lines (pronunciation practice), and study 50 vocabulary items found in the videos (a type of dictation activity). At the time of data collection, students were required to have completed 40 videos, 400 spoken lines, and 400 vocabulary items in total. Many of the students had completed much more than was required for the course. One of the hypotheses of the current study is to see if students who complete more than is required from their language learning course are actually grittier and intrinsically motivated than those who are just meeting the requirements.

Research Questions

1. Does the 10-item grit questionnaire reliably separate the participants into varying degrees of grit (construct validity)?
2. Are the grit items unidimensional?
3. Does the 10-item intrinsic motivation questionnaire reliably separate the participants into varying degrees of intrinsic motivation (construct validity)?
4. Are the intrinsic items unidimensional?
5. To what degree, if any, does grittiness correlate to intrinsic motivation measures?
6. Do gritty and intrinsically motivated students study English on their own more than their less gritty and less intrinsically motivated classmates (autonomous language learning)?

Results

The instrument and data were analyzed with Winsteps 4.0 software (Linacre, 2018) using the Rasch Rating Scale model for categorical data. The Rasch analysis consisted of person and item fit analysis, item-person maps, and Rasch principal component analysis (PCA) of item residuals (Apple, 2013; Kramer, McLean, & Martin, 2017).

Person and Item Fit (Grit)

The person data were analyzed using Winsteps to ensure the participants' responses were conforming to the model. The Rasch reliability of the person responses was approximately .59 with a separation of 1.21. With participants 1 and 18 removed from the grit analysis for careless responses, the final N-size for the grit questionnaire became (N =37). This means that the model could only stratify the person responses into one level across the construct. Table 1 shows the descriptive statistics for Grit items. Aside from the

two students removed for erratic responses, three additional students had high outfit MNSQ (>2.0). Their suspect responses were investigated and one misfitting response from each person—all different items—were removed and treated as missing. After removing these responses, the person infit and outfit fell within acceptable ranges (.5 - 1.5). The small sample size in this pilot study could be the cause of person fit issues, as even a one misfitting response can cause model fit issues.

Table 1.
Descriptive statistics for Grit Questionnaire items (N = 37)

Item	Item Description	M	SD
01	(R) New assignments or projects in my language class distract me from older assignments or projects.	2.87	0.86
02	Making mistakes in another language encourages me to study harder.	4.05	0.83
03	(R) When I am studying a language, I lose interest quickly.	3.74	0.94
04	I am a hard worker.	3.28	1.02
05	(R) When I have a language learning goal, I often choose to follow a different goal later.	2.64	0.99
06	(R) I have difficulty focusing on long term projects or goals.	2.44	0.94
07	I always accomplish my language learning goals.	3.00	0.92
08	I often do more than is required when I am studying a language.	3.46	0.85
09	When I am studying a language, I cannot be distracted from my task.	2.92	0.90
10	(R) I get disappointed and give up when I am unsuccessful.	2.85	1.23

Note. A Likert-like scale from (1) Not like me at all to (5) Exactly like me

As for item fit, the data look much more positive. The Rasch item fit analysis found that the reliability of the instrument was .93 with a separation of 3.71 (Table 2). A separation of 3.7 shows that the participants were able to distinguish three separate levels of the construct being measured by the items. All item mean-square scores and z-scores are within the appropriate criteria (0.5-1.5 MNSQ and ±3.0 for ZSTD), indicating that no items need to be removed from the questionnaire (Bond & Fox, 2015; Linacre, 2013).

Table 2
Item fit statistics for Grit Items (N = 37)

Item	Measure	S.E.	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
6	1.08	0.21	0.99	0.04	1.01	0.13
5	0.82	0.21	1.03	0.23	1.05	0.31
10	0.61	0.25	1.40	1.72	1.37	1.61
1	0.43	0.21	0.79	-0.98	0.79	-1.01
9	0.39	0.21	1.08	0.44	1.09	0.47
7	0.26	0.21	0.70	-1.48	0.69	-1.57
4	-0.09	0.21	0.88	-0.48	0.86	-0.58
8	-0.55	0.22	0.92	-0.27	0.95	-0.16
3	-1.12	0.24	0.92	-0.28	0.92	-0.25
2	-1.83	0.28	1.23	0.97	1.22	0.96

Note. Measure in Rasch logits, S.E. = standard error, MNSQ = mean squared, ZSTD = standard z-scores.

Additionally, Figure 1 shows the person-item map in which person ability is indicated by *X* marks on the left and item difficulty is measured from low to high on the right. The higher the person on the map, the grittier they are; and the higher the item on the right, the more difficult it is to endorse. As can be seen on the map in Figure 1 and in the item fit statistics in Table 2, the perseverance sub-set of the grit construct were the for most difficult to endorse items. These items were also the items with reverse valances.

A Rasch Principle Components Analysis (PCA) of the standard residuals was conducted to check the unidimensionality of the grit construct. Results indicated that 41.8% of the variance (eigenvalue = 7.19)

was explained by the grit construct and the principal contrast explained 11.1% of the variance (eigenvalue = 1.92). According to dimensionality guidelines Linacre (2013), if the unexplained variance explained by the 1st contrast has an eigenvalue less than 2.0, the possible second dimension has fewer than 2 items and is not likely an issue. These results suggest that that grit construct here is probably unidimensional, but some items might be problematic. Table 3 shows the results of the Rasch PCA. Analysis of the items with negative loadings show that most of these items are "perseverance" sub-scale items with reversed valances. Items 8 and 4 address diligence or passion to complete tasks or do more than is necessary, whereas Items 1 and 9 address inability to focus or distractibility. These two facets seem to be complimentary and might not represent separate constructs. These results, while tentative and sample-dependent, do not support Credé's (2018) suggestion that the grit variable is actually comprised of two independent constructs, passion and perseverance. A larger, more varied sample and a confirmatory factor analysis would be required to definitively contradict the claims made by Credé (2018). That being said, the results of the PCA suggest that the grit instrument is acting unidimensionally and person ability measurements from the Rasch analysis can be used in further analyses.

```
MEASURE PERSON - MAP - ITEM
        <more>|<rare>
    2          +
               |
               |T
           X T|
               |
         XXXX  |
    1      X  +  (R)I have difficulty focusing on long term projects or study goals
           XX S|S (R)When I have a language learning goal, I often choose to follow a different goal later
          XXX  |  (R)I get disappointed and give up when I am unsuccessful
           XX  |  (R)New assignments in my language classes sometimes distract me from older assignments
          XXX  |  I always accomplish my language learning goals
               |  When I am studying a language, I cannot be distracted from my task
      XXXXXXX M|
    0     XXX  +M
           XX  |  I am a hard worker
          XXX  |
            X  |  I often do more than is required when I am studying a language
              S|
            X  |S
   -1       X  +
           XX  |  (R)When I am studying a language, I lose interest quickly
            X T|
               |
               |T
               |  Making mistakes in another langauge encourages me to study harder
   -2          +
         <less>|<freq>
```

Figure 1. Grit item-person map

Table 3
Principle Component Analysis of Standard Residuals for Grit Construct

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
8	**.61**	-0.55	0.92	0.95
4	**.56**	-0.09	0.88	0.86
3	**.51**	-1.12	0.92	0.92
7	.27	0.26	0.70	0.69
2	.20	-1.83	1.23	1.22
1	**-.58**	0.43	0.79	0.79
9	**-.55**	0.39	1.08	1.09
5	-.34	0.82	1.03	1.05
10	-.25	0.61	1.40	1.37
6	-.18	1.08	0.99	1.01

Note. Measure is in Rasch logits. Positive loadings indicate the items likely measure the intended construct whereas negative loadings indicate a possible subdimension.

Person and Item Fit (Intrinsic Motivation)

A Rasch analysis was conducted with Winsteps 4.0 to assess the reliability and construct validity of the intrinsic motivation items. The participants belong to an international studies course and must study multiple foreign languages. Additionally they are required to study abroad in a country of their choosing in the 2nd or 3rd year in university. It is not surprising that the mean responses for each of these intrinsic motivation items would be so high (Table 4). With participants 1 and 18 removed, person reliability for the intrinsic motivation items was .77 with a person separation of 1.81. These results suggest a targeting problem in the items. Furthermore, the Rasch analysis showed that the item reliability of the intrinsic items was .90 with an item separation of 2.93. These results suggest that the people were able to identify at least two different levels in the intrinsic motivation construct. Similar to the grit instrument, some of the participants gave unexpected responses. One participants' response to one item was removed from the study and treated as missing. Removing this response did not significantly change the reliability measurements or the item fit statistics found in Table 5.

Table 4
Descriptive statistics for Intrinsic Motivation Questionnaire items (N = 37)

Item	Item Description	*M*	*SD*
1	I study languages to improve myself.	4.33	0.66
2	I enjoy learning new things in another language.	4.26	0.75
3	I get pleasure from using another language.	4.15	0.81
4	I enjoy the challenge of trying to learn another language.	4.15	0.74
5	I feel excited when I can use something I have learned recently.	4.46	0.76
6	I want to be someone who can speak more than one language.	4.69	0.69
7	I study other languages in order to understand the culture better.	3.77	1.01
8	I often feel that studying a language will help me in the future.	4.82	0.45
9	I enjoy interacting with people in other languages.	4.36	0.78
10	I feel satisfied when I complete challenging activities in a foreign language.	4.33	0.77

Note. A Likert-like scale from (1) Not like me at all to (5) Exactly like me

Looking at the item fit statistics (Table 5) and the person-item map (Figure 2), it is fairly clear that the items did not target these particular participants' intrinsic motivation well. Most of the participants were able to able to easily endorse even the most difficult item, *I study other languages in order to understand*

the culture better. This is a problem of instrument design and targeting with this specific group of students. It is not recommended to make general conclusions of item fit for such a small sample size.

Table 5
Item Fit Statistics for Intrinsic Items (N = 37)

Item	Measure	*S.E.*	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Outfit ZSTD
7	1.58	0.31	1.29	1.18	1.27	1.12
3	0.83	0.29	0.62	-1.63	0.59	-1.82
4	0.74	0.29	0.78	-0.83	0.82	-0.65
2	0.49	0.30	0.75	-0.95	0.68	-1.29
1	0.22	0.30	0.82	-0.67	1.00	0.09
10	0.22	0.36	1.44	1.58	1.38	1.33
9	0.13	0.31	0.96	-0.06	0.97	-0.03
5	-0.27	0.35	1.21	0.87	1.09	0.39
6	-1.78	0.48	1.34	1.15	1.02	0.22
8	-2.15	0.48	1.10	0.39	0.90	0.08

Note. Measure in Rasch logits. S.E. = standard error, MNSQ = mean squared, ZSTD = standard z-scores.

A principle components analysis conducted on the intrinsic items showed that the measures explained 49.7% of the variance with an eigenvalue of 9.86, while the first contrast explained 10.5% of the unexplained variance (eigenvalue = 2.08). Table 6 shows the loadings for intrinsic items. These results indicate that the intrinsic items explain a significant amount of the variance and is likely unidimensional. Item 7, *I study other languages in order to understand the culture better,* and Item 4, *I enjoy the challenge of trying to learn another language*, might represent different aspects of motivation, for instance, instrumental or extrinsic rather than intrinsic motivation since these two items are dealing with purposes for learning rather than feelings about learning. However, when comparing these items to Item 10 and Item 8, there does not seem to be a significant deviation in theme. They seem to describe satisfaction gained by studying and perceived utility of studying a foreign language.

Table 6
Principle Component Analysis of Standard Residuals for Intrinsic Motivation Construct

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
10	**.83**	0.22	1.44	1.38
8	**.63**	-2.15	1.10	0.90
5	.14	-0.27	1.21	1.09
9	.13	0.13	0.96	0.97
3	.09	0.83	0.62	0.59
1	.04	0.22	0.82	1.00
7	**-.74**	1.58	1.29	1.27
4	**-.57**	0.74	0.78	0.82
2	-.25	0.49	0.75	0.68
6	-.04	-1.78	1.34	1.02

Note. Measure is in Rasch logits. Positive loadings indicate the items likely measure the intended construct whereas negative loadings indicate the likely presence of a secondary dimension.

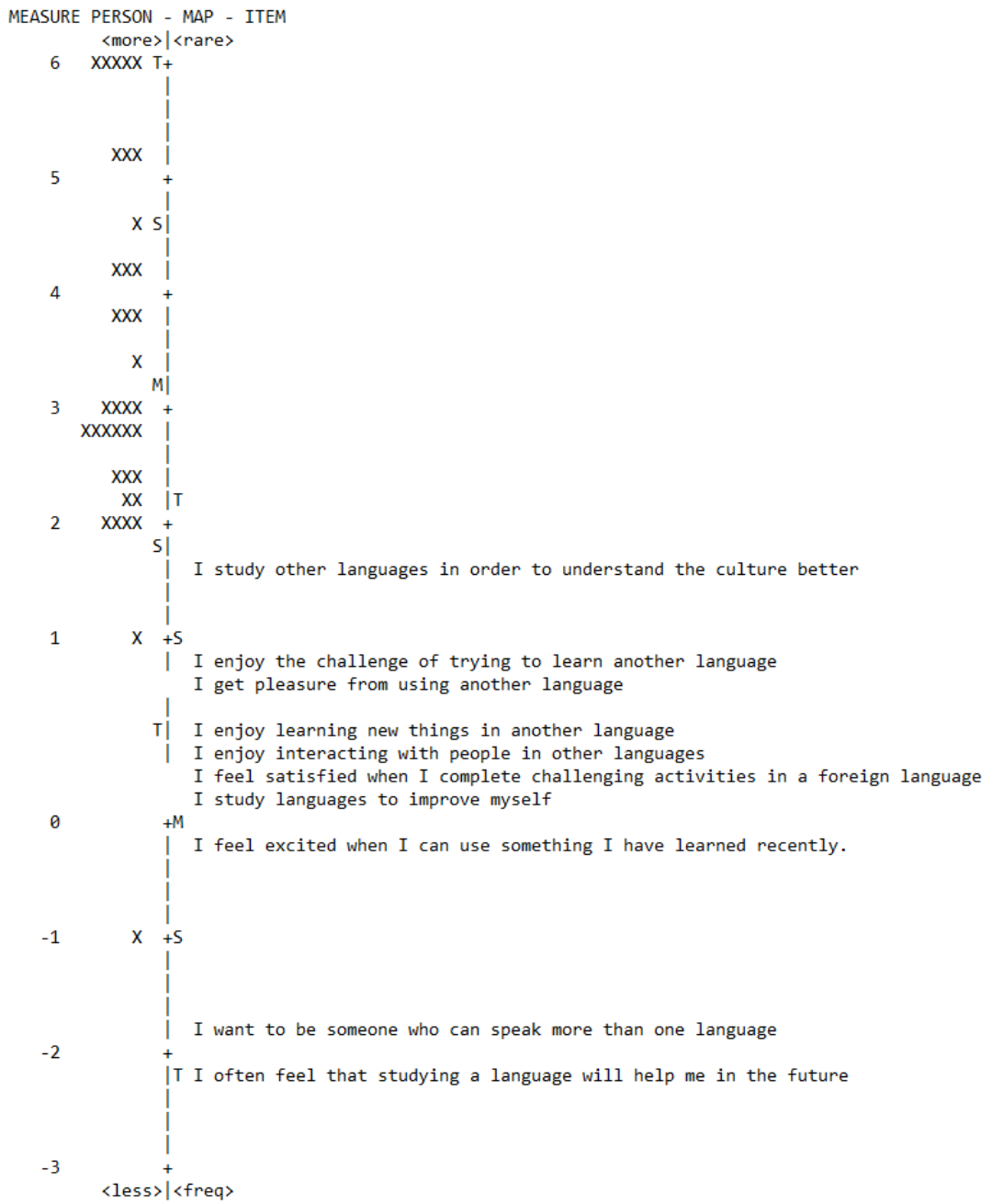

Figure 2. Intrinsic motivation item-person map

Correlation Analysis

In order to see if grit and intrinsic motivation are connected in any way, the person logit measurements for both constructs were exported into SPSS and a correlation analysis was conducted. Pearson correlation coefficient suggested that the two constructs are moderately positively related (r = .40, p = .014). Cohen (1988) offers general psychological effect size interpretation guidelines which suggest that a significant Pearson correlation between r =.30-.50 be classified as having a medium effect. A more recent set of

guidelines for interpreting correlation effect sizes has been suggested by Plonsky and Oswald (2014). In their meta-analysis, they suggest that *r* correlation statistics in the second language research should be amended to represent their actual distribution in the research. They conclude that in L2 research, "*r*s close to .25 [should] be considered small, .40 medium, and .60 large" (p. 889). The Pearson correlation coefficient of .40 suggests that 16% of the variance in the measurement is explained by the relationship between grit and intrinsic motivation, and by either standard, this relationship is one of medium effect.

However, this correlation might not tell the whole story due to the low person reliability of the grit instrument. The use of disattenuated correlations can show if the correlation between two measurements is a result of measurement error in the instrument or if the two scores are not actually correlated. Following a formula to calculate disattenuated correlations from a Pearson correlation (Statistica Help, n.d.), the disattenuated correlation between grit and intrinsic motivation showed that the two constructs are highly correlated when taking error into account ($r = .59$). These results should be taken with a grain of salt, and are not to be used to assess the correlation between the two measures. Disattenuation is presented here as a suggestion to future researchers interested in grit and motivation. The low person reliability of these instruments in this pilot study might be causing an artificially smaller relationship.

The next step in determining the value of these two measurements is to see if they correlate to student effort. Data regarding participants' English Central usage was plotted in a correlation analysis to see if the participants who are grittier or more highly intrinsically motivated actually study more on the app. Participants in the current study were required to watch 40 videos, study 400 vocabulary words, and speak 400 lines at the time of data collection. Some students completed noticeably more than the required amount while others did not use the program at all. Table 7 shows the descriptive statistics for English Central video usage, vocabulary words studied, and lines spoken.

Table 7
Descriptive Statistics for English Central Usage

	N	Min	Max	*M*	*SD*	Skew	SES	Kurt	*KES*
Videos	37	0.00	104	41.55	22.85	0.75	0.38	1.36	0.75
Vocabulary	37	0.00	810	357.05	191.75	0.08	0.38	0.25	0.75
Lines	37	0.00	927	345.63	202.55	0.39	0.38	1.07	0.75

Note. Required amount was 40 videos, 400 vocabulary, and 400 spoken lines.

Spearman's rank correlation coefficients (r_s) were calculated for grit, intrinsic motivation, and the three English Central usage variables because it is more robust to violations of normal distribution and combinations of variable type, in this case continuous and ordinal count data (Larson-Hall, 2016). The results in Table 8 show that grit and intrinsic motivation are moderately correlated at $r_s = .46$, $p = .004$. The three English Central usage counts were all highly correlated with each other (all greater than $r_s > .70$, $p < .001$), as was to be expected. None of the EC measures correlated significantly with the grit measures. However, intrinsic motivation was significantly correlated to one EC variable, spoken lines ($r_s = .39$, $p = .016$). This suggests that participants with higher intrinsic motivation are more inclined to practice speaking using the English Central application.

In order to get a clearer view of the relationships between the five variables, they were all investigated using a scatterplot matrix (Figure 3). Using this figure, the relationship between grit and intrinsic motivation seems to be following a central, upward tendency. Additionally, all three counts of English Central data show a clear, positive relationship as noted in Table 8. Finally, the relationship between intrinsic motivation and the number of lines spoken is less apparent.

Table 8
Spearman's Rho Correlation Coefficients for Grit, Intrinsic Motivation, and English Central Usage

	Intrinsic	Videos	Vocab	Lines
Grit	.46*	.15	.23	.26
Intrinsic		.30	.12	.39*
Videos			.71**	.74**
Vocabulary				.74**

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

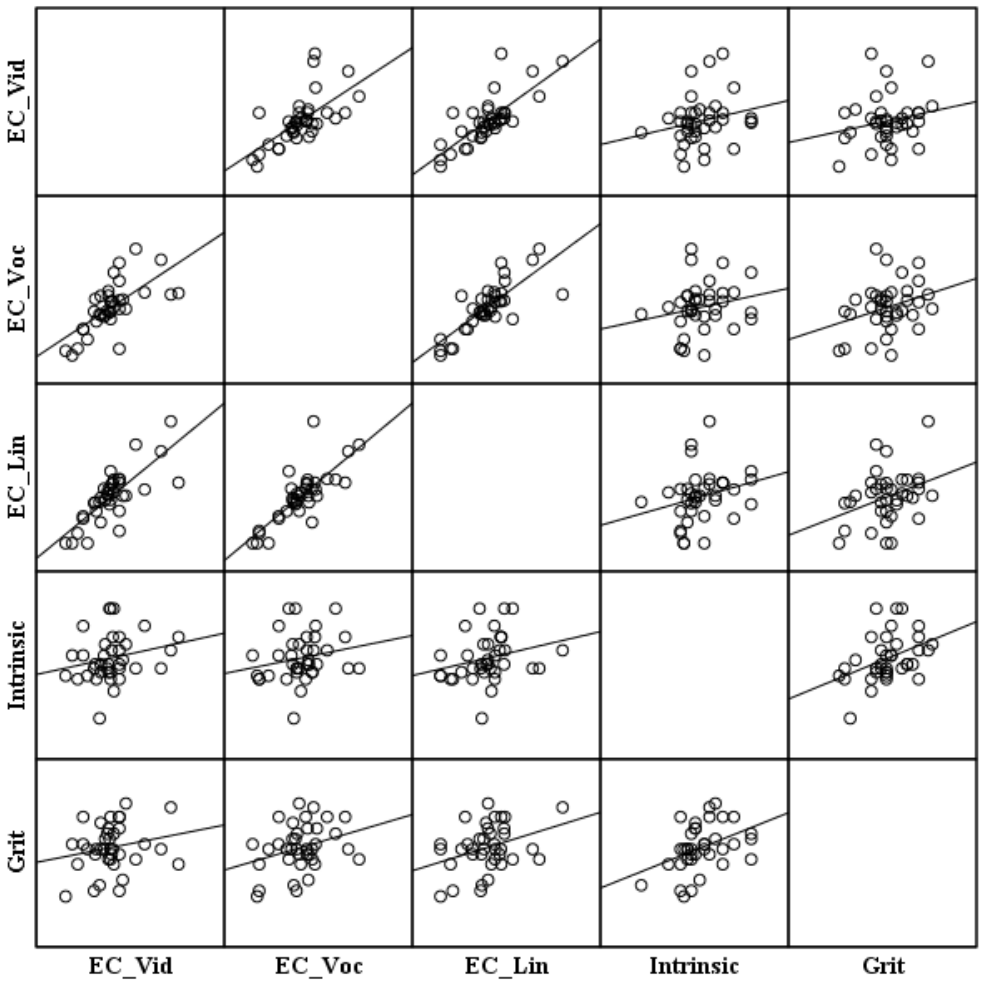

Figure 3. Scatterplot of grit, intrinsic motivation, and EC variables

Discussion

It is necessary to discuss each of the research questions one-by one. The first research question was designed to test one aspect of construct validity of the grit instrument. Do the 10 items adapted from Duckworth and associates reliably separate the participants into varying degrees of language learning grit? The answer seems to be yes and no. The person reliability of the responses was approximately .59 with a

separation of 1.21, yet the item reliability was .93 with a separation of 3.71. Seemingly the participants were able to distinguish between at least three different levels of the construct, but the instrument is not sensitive enough to separate the people as well. This may be a problem of item wording. Clearly some items were more difficult to endorse than others, as Figure 1 shows. However, the low person reliability of instrument is an issue with instrument design. Increasing the number of unidimensional and well-targeted items would likely increase the person reliability on future versions of this instrument as would increasing sample size by surveying individuals with various degrees of grit.

The second research question referred to dimensionality. Does this questionnaire address one construct, grit? The PCA seems to suggest that the grit instrument is unidimensional but some of the items suggest a small subdimension accounting for approximately 25% of the variance. This is not significant enough to suggest that grit is actually two separate constructs. Passion and perseverance for long term goals could actually be two separate but connected constructs, however the limitations of the current pilot study prevent further investigation. Factor analysis would be a better methodology to address Credé (2018) suggestion that grit is two constructs. Within the current study, it is possible that the negative valance of the items distracted participants, resulting in a noticeable subdimension in the instrument. A social desirability threat to validity could also have an effect on the responses.

The third research question, do the ten intrinsic motivation items reliably separate the participants into varying degrees of intrinsic motivation, did not yield conclusive results. Even though the intrinsic items were adapted from previous intrinsic motivation research, there was not enough spread to successfully separate the participants into varying motivation levels. Most of the items were far too easy to endorse by these participants. This is likely due to the participants surveyed in this study. As previously mentioned, they are all international studies students. The majority of them are women and they are all currently preparing to study abroad next year. They are all highly motivated students. More difficult items or a wider variety of participants would provide better spread of person ability and item targeting.

Research question four dealt with intrinsic items' unidimensionality. The results of the PCA showed that the intrinsic items were unidimensional with the exception of one or two items which may not be tapping the intrinsic motivation construct. Deleting Item 7 and Item 4 from the intrinsic motivation instrument reduced item reliability to .88 and did not change the explained variance in the PCA. Leaving those items in the questionnaire does not seem to have a detrimental effect on unidimensionality or reliability.

Addressing the fifth research question, the test of relationship between grit and intrinsic motivation resulted in a significant positive correlation of medium effect (r = .40). Both grit and motivation have an aspects of perseverance, passion, and stick-to-itiveness. It seemed natural at the outset of this investigation that these two constructs would be significantly correlated. As the results showed, grit and intrinsic motivation are highly correlated with a disattenuated Pearson correlation of r = .59. There are generalizability concerns with this study and as such the results should be interpreted carefully. The participants in this study might not be representative of the population. Further research is required.

Lastly, do these construct measurements correlate to a real-world, autonomous language learning measurement? When comparing the person grit and intrinsic logit measurements to English Central usage statistics, there did not seem to be any significant relationship. Increasing person reliability, as discussed above, would rectify this discrepancy. It would allow for a more direct comparison without the need of disattenuation. Alternatively, it is also possible that English Central homework counts are not a valid proxy measurement of grit or motivation. This is a required homework assignment. I had hoped that the wide range in achievement (see min, max, and mean in Table 7) for the English Central data would provide enough variance in the data to overcome the fact that these tasks were required by the curriculum and not completed out of desire to study only.

Limitations

There are many limitations to the current pilot study. First, the participant sample and size was a matter of convenience. If the instrument were to be given to larger sample with varying degrees of motivation and grit, the results might vary. The lack of representation of males in the study is also an issue to be addressed. Additionally, these students are all international studies majors. They all intend to study abroad next year and many have already had study abroad experiences. These experiences may color the participants' motivation responses.

Finally, the grit questionnaire contained items with negative valance. Research and experience tends to agree that people do not want to answer negatively about themselves. Kramer, McLean, and Martin (2017) found that even when translated into Japanese, these items were problematic. The current study used only English items. Future grit research cannot continue to use these psychometrically awkward items.

Conclusion

In conclusion, the current study explored the relationship of language learning grit and intrinsic motivation to learn a language while at the same time validating two questionnaires. The lack of research on grit in the second language learning makes this exploration worthwhile. However, the similarity of the grit construct to other more well-established personality constructs like conscientiousness seem to provide doubt regarding the value of the construct at all (Credé, Tynan, & Harms, 2017; Credé, 2018; Rimfeld, Kovas, Dale, & Plomin, 2016). The conclusion of this pilot study—grit and intrinsic motivation are moderately correlated—suggest that with improvements in methodology, item design, and targeting, the grit construct could be a valuable addition to individual differences toolbox for L2 researchers.

References

Apple, M. T. (2013). Using Rasch analysis to create and evaluate a measurement instrument for foreign language classroom speaking anxiety. *JALT Journal*, *35*(1), 5-28.

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

Cohen, J. (1988). *Statistical power analysis for behavioral sciences* (2nd ed.). New York, NY: Lawrence Erlbaum Associates.

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology, 113*(3), 492-511. DOI:10.1037/pspp0000102

Credé, M. (2018). What shall we do about grit? A critical review of what we know and what we don't know. *Educational Researcher, 47*(9), 606-611. DOI:10.3102/0013189X18801322

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York, NY: Plenum.

DeKeyser, R. (2007). Introduction: Situating the concept of practice. In: DeKeyser, R. (ed.) *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge, UK: Cambridge University Press.

Duckworth, A. L., & Eskrels-Winkler, L. (2013). True Grit. *Observer, 26*(4). Retrieved from https://www.psychologicalscience.org/observer/true-grit

Duckworth, A. L., Peterson, C., Matthews, M., & Kelly, D. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*(6), 1087-1101. DOI: 10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the short grit scale (Grit-S). *Journal of Personality Assessment, 91*(*2*), 166-174. DOI:10.1080/00223890802634290

English Central. (n.d.). About English Central. Retrieved from https://www.englishcentral.com/static/corporate/section/

Kramer, B., McLean, S., & Martin, E. S. (2017). Student grittiness: A pilot study investigating scholarly persistence in EFL classrooms. *大阪女学院短期大学紀要第, 47,* 25-41.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R* (2nd ed.). New York, NY: Routledge.

Linacre, J. M. (2018). WINSTEPS (Version 4.0). Retrieved from http://www.winsteps.com/

Linacre, J. M. (2013). *A user's guide to WINSTEPS*. Chicago, IL: MESA. Retrieved from http://www.winsteps.com/

Noels, Clément, and Pelletier (2001). Intrinsic, extrinsic, and integrative orientations of French Canadian learners of English. *The Canadian Modern Language Review, 53*(3), 424-442. DOI: 10.3138/cmlr.57.3.424

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

Rimfeld, K., Kovas, Y., Dale, P. S., & Plomin, R. (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of personality and social psychology*, *111*(5), 780-789. DOI: 10.1037/pspp0000089

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68-78. DOI:10.1037/0003-066X.55.1.68

Schmitt, N., & Stults, D. M. (1985). Factors defined by negatively keyed items: The result of careless respondents? *Applied Psychological Measurement*, *9*, 367–373. DOI:10.1177/014662168500900405

Statistica Help (n.d). *Reliability and item analysis introductory overview – Correction for attenuation.* Retrieved from http://documentation.statsoft.com/STATISTICAHelp.aspx?path=Reliability/ReliabilityandItemAnalysis/Overview/ReliabilityandItemAnalysisIntroductoryOverviewCorrectionforAttenuation

Swain, S. D., Weathers, D., & Niedrich, R. W. (2008). Assessing three sources of misresponse to reversed Likert items. *Journal of Marketing Research*, *45*, 116–131. DOI:10.1509/jmkr.45.1.116

Appendix A: Construct Maps

Construct Map:

Grit for Autonomous Language Learning

High Grit	+1	• (R) When I have a language learning goal, I often choose to follow a different goal later. • (R) I get disappointed and give up when I am unsuccessful. • When I am studying a language, I cannot be distracted from my task. • (R) New assignments or projects in my language class distract me. from older assignments or projects. • I always accomplish my language learning goals • I am a hard worker • I often do more than is required when I am studying a language • (R)When I am studying a language, I lose interest quickly.
Moderate Grit	0	• Making mistakes in another language encourages me to study harder
Low Grit	-1	

Note. (R) means the valance of the items is reversed. These codes were reversed at data entry.

Construct Map:

Intrinsic Motivation for Autonomous Language Learning

High Intrinsic	+1	• I study other languages in order to understand the culture better. • I enjoy the challenge of trying to learn another language. • I get pleasure from using another language. • I enjoy learning new things in another language. • I enjoy interacting with people in other languages.
Moderate Intrinsic	0	• I feel satisfied when I complete challenging activities in a foreign language. • I study languages to improve myself. • I feel excited when I can use something I have learned recently. • I want to be someone who can speak more than one language.
Low Intrinsic	-1	

Appendix B: Test Specifications

Test Specification Table	
Time allowed	As much time as needed; approximately 10 minutes
Test Delivery	Google Forms; Delivered via link on participants' learning management system
Skill Focus	Self-reflection; ability to read statements and discern the degree with which the statements describe the participants
Task Description	Read a short statement or situation and decide how much the situation describes their own opinion
Language level	Statement language should contain frequent words that do not require a dictionary; language should fall within the first 2k vocabulary band of English on the BNC
Expected response	Exactly like me (5); Mostly like me (4); Somewhat like me (3); Not much like me (2); Not like me at all (1)
Discourse purpose	To discover the participants' level of grit and intrinsic motivation for language learning.
Scoring Parameters	Likert-like; degree to which the statement describes their opinion; Some items require score reversal. Scores will be reversed upon data entry.
Instructions to candidates	Here are a number of statements that may or may not apply to you concerning your language learning experiences. Think about how you compare to most language learners in the world, not just your friends and classmates. There are no wrong answers, just answer honestly.
Guiding language	• "I" Statements are acceptable. • Avoid verbs with negative connotations when possible • Reverse coded questions should be careful not to imply negatively or bias the students against choosing the item.. No one wants to say that "I give up easily" is very much like themselves • Avoid passive and complicated grammatical structures, L2 learners may not be able to understand them. • Do not include the scores for the items next to the choices, this may bias the participants' choices.
Additional details	Participants can provide additional comments and questions regarding the questionnaire in English or Japanese. Students may also sign up to participate in a follow-up interview regarding their responses

Questions and answers about language testing statistics:

Overall English proficiency (whatever that is)

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

As more and more language tests are developed, typically language teachers often want to know how the scores on new tests relate to more familiar tests scores. This seemly particularly true among tests which claim to measure *general English ability*. … can we say with confidence that there is, in fact, such a thing as general English ability?

Answer:

This is a question that I have been wrestling with for my whole career in language testing, and now, you seem to be having doubts about it too. What you are referring to as *general English ability* is also sometimes called *overall English proficiency* (or ELP), which is how I will refer to that idea here. From as far back as 1977, whenever I have said the words *overall English proficiency*, I have added (soto voce) *whatever that is*, which comes out something like "overall English proficiency (whatever that is)." Recently, I've been working on a number of papers circling this issue, but now I'm writing one that focuses directly on this topic. As a result, I have been doing a fair amount of thinking about the issue. Let me share some of my preliminary thoughts with you now in the hope that they will help answer your question and entice you into later tracking down the larger paper that I will eventually publish.

I think the central issue involved in the fuzziness of the overall ELP concept has to do with something else I have been dealing with for over four decades. Often when I say I'm specialized in *language testing*, teachers and researchers in other areas of the field ask me why language testing is so far behind the rest of the field (by which I think they mean *why are overall ELP tests like the TOEFL, TOEIC, IELTS, etc. so far behind the rest of the language teaching field*). I have typically answered rather defensively/snidely that it is much harder to operationalize (that is measure quantitatively) the many variables in our field than it is to sit around in an armchair and think them up. But my view of all that has changed now that I am retired and sitting around in an armchair way too old to be defensive. The bottom line is that developments in *language teaching* have exploded during my time in the field, and far outpaced changes in the overall ELP tests, which are trailing far behind. There are at least three main areas of change:

- Expansion of our views on the *nature of language learning*
- Growth in the number of *pedagogical options* available to teachers
- Opening up of our ideas about *who owns English*

Let me take each of those sets of issues separately.

Nature of Language Learning

As mentioned earlier, our conception of the *nature of language learning* has expanded enormously. Language testers have long discussed ways that our conceptions of the nature of language learning have expanded in the following stages that built one on the other:

1. language knowledge (Lado, 1961)
2. linguistic knowledge vs. channel control (Carrol, 1961)
3. competence vs. performance (Chomsky, 1965)
4. separate scales for skills with well-educated NS at top (ILR, 1985)
5. multiple dimension model of communicative competence (Canale & Swain, 1979, 1980, 1981, and Bachman, 1990)

Thus, from primitive views of language knowledge (e.g., where language learning involves gaining knowledge of phonemes, vocabulary and grammar ["1" above]) through the distinctions between linguistic knowledge vs. channel control ["2" above] and competence vs. performance ["3" above], we have managed to assess passive knowledge (green) (mostly the knowledge components to the left in each case) to the exclusion of active knowledge. Later more complex views of language competence developed that separated ELP into skills ["4" above] and a multidimensional model of communicative competence ["5" above], each of which are currently measured only partially (yellow).

The multidimensional model of communicative competence can be outlined as follows:

1. **organizational competence**
a. *grammatical* (i.e., vocabulary, syntax, morphology, & phonology/graphemes)
b. *textual* (i.e., cohesion & rhetorical)
2. **pragmatic competence**
a. *illocutionary* (i.e., ideational, heuristic, manipulative, & imaginative functions)
b. *sociolinguistic* (i.e., differences in dialect/variety, naturalness, differences in register, & cultural references and figures of speech)

Notice in the multidimensional model, that only a small portion is shaded in green. Perhaps this should be green and yellow because these aspects of ELP are largely measured passively and only partially in the sense that the spoken grammar, phonology of connected speech, etc. are not assessed at all. Note also that the *textual* subpart of organizational competence and the entire pragmatics competence second half of the model are typically not represented at all in the overall ELP tests.

Pedagogical Options

One set of *pedagogical options* in education is generally outlined in the taxonomy of the cognitive domain (Krathwohl, 2002):

1.0 Remember (Retrieving relevant knowledge from long-term memory)
  1.1 Recognizing
  1.2 Recalling
2.0 Understand (Determining the meaning of instructional messages, including oral, written, and graphic communication)
  2.1 Interpreting
  2.2 Exemplifying
  2.3 Classifying
  2.4 Summarizing
  2.5 Inferring
  2.6 Comparing
  2.7 Explaining
3.0 Apply (Carrying out or using a procedure in a given situation)
  3.1 Executing
  3.2 Implementing

4.0 Analyze (Breaking material into its constituent parts and detecting how the parts relate to one another and to an overall structure or purpose)
    4.1 Differentiating
    4.2 Organizing
    4.3 Attributing
5.0 Evaluate (Making judgments based on criteria and standards)
    5.1 Checking
    5.2 Critiquing
6.0 Create (Putting elements together to form a novel, coherent whole or make an original product)
    6.1 Generating
    6.2 Planning
6.3 Producing

Notice that the lowest levels cognitive skills of 1.0 remembering and the interpreting part of 2.0 understanding are typically tested on the overall ELP exams, at least passively, but that the rest of the middle to higher order skills are not involved at all. These lowest levels may cover the skills that undergraduate students are required to be able to do in English prior to heading for university, but they are not at all sufficient for the English needed to perform all the middle to higher order skills shown in 2.2 to 6.3 that graduate students (who also take these tests for university admissions) are required to be able to do in English.

Over the years, pedagogical approaches in language teaching have multiplied greatly**.** Consider the following approaches which each represent different belief systems that language teachers may hold singly or in combination (After Brown, 2016):

1. Classical Approach
2. Grammar Translation Approach
3. Direct Method Approach
4. Audiolingual Approach
5. Cognitive Approach
6. Communicative Approach

Notice that five of the six pedagogical belief systems only passively or partially underlie the typical overall ELP tests.

Consider also the following syllabuses which each represent different basic units around which language teaching/curricula are typically organized (after Brown 2016):

1. Structural
2. Situational
3. Topical
4. Functional
5. Notional
6. Skills-based
7. Task-based
8. Lexical
9. Pragmatic
10. Genre-based
11. Discourse-based
12. Communicative strategies

Notice that one of these syllabuses is typically covered typically covered in terms of passive knowledge by the overall ELP tests, while three can be said to be partially covered. That leaves seven syllabuses that are not usually covered at all. Think about the message that sends to teachers and students about the importance of these various syllabuses.

Who owns English?

Another way that the field has changed in recent years is in our ideas about *who owns English.* Essentially, we have moved from firmly believing in the native-speaker model of English (meaning that it is owned by native speakers) to the recognition of World Englishes, including Inner Circle, Outer Circle, and Expanding Circle Englishes (Kachru, 1986) as shown below:

World Englishes
Inner Circle **(**UK English**,** North American English, Australian/New Zealand
Outer Circle (e.g., Singaporean English**,** Indian English**,** Jamaican English, etc.)
Expanding Circle (e.g., German English**,** Chinese English**,** Arabic English, etc.)

I have shaded the inner-circle Englishes in yellow because even they are typically partially represented in an idealized sort of educated or broadcast English, which of course ignores the reality of the huge variations in English due to class, education, location, and dialect that exist in each of these native-speaker Englishes. The overall ELP examinations typically ignore the outer and expanding circle Englishes altogether, saying things like our examinees are being admitted to North American universities so that is the English they will need. Never mind that the Japanese engineering graduate student admitted to a university based on her TOEFL score is far more likely to interact with outer and expanding circle speaking students and professors in her day-to-day life than with inner-circle native speakers—and even those few native speakers are likely to use all sorts of dialects that are definitely not the idealized North American sort of English.

Conclusion

In this column, I have tried to show how the expansion of our views on the *nature of language learning*, the growth in the number of *pedagogical options* available to teachers, and the opening up of our ideas about *who owns English* have outpaced any changes in the overall ELP tests that we use for important decisions about admissions to university (for undergraduate and graduate students alike). In direct answer to your question, I believe that the overall ELP examinations by and large assess general English ability (or overall ELP) only partially and most of that is focused on passive knowledge. Taking into account even the small number of the issues discussed in this column, it is hard to conclude that the overall ELP tests are adequately measuring overall ELP.

Worse yet, depending on which limited set of these many aspects of ELP particular tests decide to include in their design, the so-called overall ELP tests may be assessing quite different things and therefore may ***not*** be directly comparable despite the fancy tables that have been produced by various organizations.

I hope this column addressed your question(s) adequately and provided you with the information you need for thinking about *general English ability*, or if you prefer, *overall English proficiency*.

References

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University.

Brown, J. D. (2016). *Introducing needs analysis and English for specific purposes.* New York: Routledge.

Brown, J. D. (2016). *Statistics Corner: Questions and answers about language testing statistics.* Tokyo: JALT TEVAL SIG.

Canale, M., & M. Swain. (1979). Testing and communicative competence. Paper presented at the 13th annual TESOL Convention, Boston.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics, 1*, 1-47.

Canale, M., & Swain, M. (1981). A theoretical framework for communicative competence. In A. S. Palmer, P. J. M. Groot, & G. A. Trosper (Eds.) *The construct validation of tests of communicative competence* (pp. 31-36). Washington, DC: TESOL.

Carrol, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp. 30-40). Washington, DC: Center for Applied Linguistics.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT.

ILR. (1985). *Language skill level descriptions* (Internal document). Washington, DC: Interagency Language Roundtable.

Kachru, B. B. (1986). The power and politics of English. *World Englishes, S2/3*, 121-140.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212-218.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.

A brief coda

All good things must come to an end, and thus, ends my contribution to this column. I have had the honor of writing this column two or three times per year for a little over 22 years (since Issue 1 Number 1 of the *Shiken* in April 1997). It has always been an interesting process, much of which was included in the book published by the JALT TEVAL (see Brown, 2016, in the references above). None of this would have been possible without the many questions submitted by readers, JALT TEVAL officers and members, Shiken editors, and the many graduate students who have passed through my classes on their way to better things. I thank them all for their curiosity and support, and I wish them all the very best in the years to come.

Aloha nui loa, JD

Professor JD Brown, retired
Department of Second Language Studies
University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822 USA

Call for Papers

Shiken is seeking submissions for publication in the June 2020 issue. Submissions received by 1 January, 2020 will be considered, although earlier submission is encouraged to allow time for review and revision. *Shiken* aims to publish articles concerning language assessment issues relevant to classroom practitioners and language program administrators. This includes, but is not limited to, research papers, replication studies, review articles, informed opinion pieces, technical advice articles, and qualitative descriptions of classroom testing issues. Article length should reflect the purpose of the article. Short, focused articles that are accessible to non-specialists are preferred and we reserve the right to edit submissions for relevance and length. Research papers should range from 4000 to 8000 words, but longer articles are acceptable provided they are clearly focused and relevant. Novice researchers are encouraged to submit, but should aim for short papers that address a single research question. Longer articles will generally only be accepted from established researchers with publication experience. Opinion pieces should be of 3000 words or less and focus on a single main issue. Many aspects of language testing draw justified criticism and we welcome articles critical of existing practices, but authors must provide evidence to support any empirical claims made. Isolated anecdotes or claims based on "commonsense" are not a sufficient evidential basis for publication.

Submissions should be formatted as a Microsoft Word (.doc or .docx format) using 12 point Times New Roman font, although plain text files (.txt format) without formatting are also acceptable. The page size should be set to A4, with a 2.5 cm margin. Separate sections for tables and figures should be appended to the end of the document following any appendices, using the section headings “Tables” and “Figures”. Tables and figures should be numbered and titled following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition*. Within the body of the text, indicate approximately where each table or figure should appear by typing “Insert Table x” or “Insert Figure x” centered on a new line, with “x” replaced by the number of the table or figure.

The body text should be left justified, with single spacing for the text within a paragraph. Each paragraph should be separated by a double line space, either by specifying a double line space from the Microsoft Office paragraph formatting menu, or by manually typing two carriage returns in a plain text file. Do not manually type a carriage return at the end of each line of text within a paragraph.

Each section of the paper should have a section heading, following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition*. Each section heading should be preceded by a double line space as for a regular paragraph, but followed by a single line space.

The reference section should begin on a new page immediately after the end of the body text (i.e. before any appendices, tables, and figures), with the heading “References”. Referencing should strictly follow the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition.*