The long and the short of it
Passage length, information density, and working memory in second language listening comprehension

Sarah Wayland, PhD
Lelyn Saner, PhD
Alexandra Ralph, BA
Stephen O’Connell, MA
Jared Linck, PhD
Debra Kramasz, MA
Kassandra Gynther, BA
Amber Bloomfield, PhD

Executive summary

PURPOSE
The most recent version of the Defense Language Proficiency Test (DLPT5) differs from the previous version (DLPT-IV) in many important respects, not the least of which is a greater emphasis on testing listening comprehension with authentic materials. Many people taking the new version of the test have performed differently than on the DLPT-IV. This has prompted questions about the impact of passage-based factors on comprehension for second language (L2) listeners.

A review of the scientific literature identified a number of passage-based factors likely to affect L2 listening comprehension. However, the research targeting these factors is limited. The University of Maryland Center for Advanced Study of Language (CASL) conducted three studies to explore the effects of some of these key variables on listening comprehension, including passage length, information density, and listeners’ working memory capacity, while carefully controlling for other factors like speech rate, speaker, and topic.

The studies reported here measured listening comprehension for native and nonnative English speakers using multiple-choice items designed to be like those used on the DLPT5 and recall items that required the listener to remember specific words heard in the listening passage. The goal of this research is twofold: to expand on what is currently known about the influence of these passage-based factors on listening comprehension and to clarify the role they play in listening comprehension situations similar to those found in the DLPT5.

CONCLUSIONS
CASL research supports the following conclusions:

1. **Participants’ L2 proficiency predicted their response accuracy for both recall and multiple-choice items.**

As expected, the English listening proficiency of nonnative speakers strongly predicted accuracy on recall and multiple-choice comprehension items. This result, while unsurprising, does indicate that the comprehension items in this study were sensitive to differences in the listening proficiency level of the participants. This conclusion is further bolstered by the finding that native English speakers showed...
considerably higher accuracy on both types of comprehension items than did nonnative speakers.

Neither the length nor the information density of passages consistently affected participants’ performance on multiple-choice comprehension items. The response accuracy for multiple-choice items did not differ systematically with either increasing levels of length or between medium- and high-density passages. This result was found for both nonnative and native speakers of English. These results suggest that length and information density are not important factors when L2 listening comprehension is measured with multiple-choice items.

Both length and information density affected participants’ performance on recall items. Length and information density of the passage impacted accuracy of recall for nonnative and native speakers. Participants recalled more words accurately when the listening passages were short (~1 minute) rather than long (~3 minutes), and when the passages’ information density was medium rather than high. For native speakers, an effect of density appeared only for passages that were also long. For nonnative speakers, high-density passages were more difficult than medium-density passages regardless of length. The results suggest that the information density of a passage is more likely to have an impact on the listening comprehension of native speakers when the passage is difficult due to other factors, such as being longer.

Participants with lower working memory capacity understood less than those with higher working memory capacity. In addition to exploring the effects of passage length and information density, CASL measured participants’ working memory capacity to determine the extent to which individual differences in working memory affected listening comprehension. As expected, individuals with higher working memory capacity understood the passages more accurately; this was reflected in the results for both the multiple-choice and recall items.

The role of working memory capacity diminished for the nonnative speakers when their English listening proficiency was taken into account. This suggests that working memory affects performance both on the English proficiency test (the listening comprehension portion of the Versant™ Pro Speaking test) and on the multiple-choice and recall comprehension items.

The relationship between working memory capacity and performance on multiple-choice comprehension items differed little between the varying levels of passage length and information density (Figure 1). This was also true for performance on the recall items. This indicates that the effects of length and information density (where present) cannot be entirely attributed to an increased load on working memory.

Figure 1. Performance on multiple-choice items did not consistently vary by the length or density of the passage for native speakers (NS) or nonnative speakers. The impact of working memory capacity on performance was also similar in all conditions.
RELEVANCE

Although density and length did not consistently impact accuracy on multiple-choice items, they did affect accuracy on recall items, suggesting that these factors do have an impact on L2 listening comprehension. When presenting listeners with tasks other than multiple-choice items, test developers and instructors should be aware of how the length and density of the listening passages can affect performance.

Although individual differences in working memory capacity predict performance on both recall and multiple-choice, they do not interact with either length or density in predicting performance. Thus, increasing levels of length or density do not seem to differentially impact listeners with higher or lower working memory capacity. In addition, for recall measures, L2 listening proficiency may override the impact of working memory on performance.

ENDNOTES

1 Authors’ names are listed in reverse alphabetical order. Each author made substantial contributions to this report.
3 CASL researchers conducted a pilot study as part of this work to establish an acceptable speech rate for the passages used in the length and information density studies. Detailed results of the pilot study are reported in Wayland, S. C., O’Connell, S., Linck, J., Kramasz, D., Gynther, K., Bloomfield, A., Blodgett, A., Silbert, N., & Saner, L. (2013). Second language listening comprehension: The impact of speech rate, passage length and information density final technical report (TTO 81434). College Park, MD: University of Maryland Center for Advanced Study of Language.

Corresponding Author and Reprints: Sarah Wayland, PhD, University of Maryland Center for Advanced Study of Language, (301) 226-8938, swayland@casl.umd.edu.

Funding/Support: This material is based upon work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government.

Nothing in this report is intended to be and shall not be treated or construed as an endorsement or recommendation by the University of Maryland, United States Government, or the authors of the product, process, or service that is the subject of this report. No one may use any information contained or based on this report in advertisements or promotional materials related to any company product, process, or service or in support of other commercial purposes. This report is not Releasable to the Defense Technical Information Center per DoD Directive 3200.12. The Contracting Officer’s Representative for this project is John Walker, Government Technical Director for Analysis at CASL, (301) 226-8912, jwalker@casl.umd.edu. DLIFLC’s Technical Task Order Manager for this project is Susan A. Whit, PhD, (831) 393-9279, susan.a.whit3.civ@mail.mil.

© 2013 University of Maryland. All rights reserved. August 2013
The Long and the Short of It: Passage Length and Information Density in Second Language Listening Comprehension

Final Technical Report

Sarah Wayland, PhD, Lelyn Saner, PhD, Stephen O’Connell, MA, Jared Linck, PhD, Debra Kramasz, MA, Kassandra Gynther, BA, Amber Bloomfield, PhD, Alexandra Ralph, BA

This report describes a set of three studies investigating the effects of passage length and information density on listening comprehension of English passages for native (L1) and non-native (L2) listeners (native Arabic and Spanish speakers). Study 1 investigates the impact of passage length (in syllables) on L2 listening comprehension when other text-based factors, such as speech rate, speaker gender, and topic, are controlled. Both Studies 2a and 2b investigate the effects on listening comprehension of information density (operationalized as type/token ratio) and passage length. Study 2a explores the impact of these factors on native listener comprehension, while Study 2b examines their impact with non-native listeners. Because information density and passage length were independently manipulated in Studies 2a and 2b, it was also possible to investigate interactions between these factors. In addition, Studies 2a and 2b measured participants’ working memory capacity so as to determine the extent to which individual differences in working memory played a role in the effects of passage length and information density.

- Study 1: The first study investigated the effects of passage length (in syllables) on non-native listening comprehension. Participants were native Arabic speakers with an intermediate level of English proficiency (Interagency Language Roundtable (ILR) 2/2+). Length, in number of syllables, was systematically varied for passages while the topic, speakers, information density, and speech rate (4 syllables per second) were kept constant across lengths. The range of lengths investigated extends both below and above those of the Level 2 passages used in the Defense Language Proficiency Test (DLPT5). Performance on recall items was poorer for the long length passages relative to short passages; however, there was no effect of length on accuracy of

---

1 Authorship is listed in reverse alphabetical order except for Alexandra Ralph, who joined the project towards the end. All other authors contributed substantially and at roughly equal levels to this report, each with their own particular expertise.

2 Results of Studies 1 and 2a have been reported elsewhere (Wayland et al., 2013) but are updated here with additional data, analyses, and interpretations.

3 As described in the literature review beginning on page 5, length of passage has been defined in a variety of ways across studies. Readers should understand that all references to length in the current studies correspond to passage length in syllables. Because speech rate was held constant across all passages at 4 syllables per second, length also corresponds to number of seconds.
responses to the multiple-choice items. English listening proficiency, as indicated by the listening
comprehension subscore of the Versant™ Pro Speaking test, significantly predicted performance
on both recall and multiple-choice items but did not interact with length.

- Study 2a: The second study examined the effects of information density (density of idea units,
  operationalized as type/token ratio [TTR]⁴) and passage length on listening comprehension for
  native listeners (L1 English) when other factors (e.g., speech rate) were controlled. Density and
  length were independently manipulated in the listening passages to allow for their separate effects
  and interactions to be investigated. The results for length replicated the findings from Study 1;
  there was no consistent impact of length on performance for the multiple-choice items, but length
  affected recall item performance. Density also affected performance on recall but did not
  consistently impact multiple-choice item performance. For recall items, recall item accuracy was
  lower for long passages, and the impact of length was exacerbated by high density. In addition,
  individuals with greater working memory capacity showed superior comprehension under all
  conditions of length and density, as reflected in the results for both the multiple-choice and word
  recall tasks. The impact of working memory capacity was surprisingly similar across all
  conditions, indicating that, while the effects of passage length varied depending on information
  density, the impact of these factors cannot be attributed solely to demands on working memory.

- Study 2b: The last study replicated the second study but with non-native listeners (L1 Spanish)
  rather than native listeners (L1 English). The results of this study indicate that both length and
  density affected comprehension for non-native listeners, but the pattern of the effect again
  differed for the multiple-choice comprehension task and the recall task. For the multiple-choice
  items, there was no consistent effect of length or density on performance. However, performance
  on recall items was worse when passages were longer or denser. Length and density did not
  interact as they did for native English speakers. As with the native English speakers, working
  memory capacity predicted performance across all length and density levels, but this relationship
  was attenuated when English listening proficiency was taken into account.

Results across the three studies indicate that the impact of passage length and information density on
listening comprehension will depend on the listeners’ task. For multiple-choice comprehension items,
length and density do not consistently affect response accuracy for either native or non-native listeners.
However, performance on recall items is affected by both length and density of the listening passage.
Further, although these factors affect recall performance for both native and non-native listeners, density
has an effect for native listeners only when a passage’s difficulty has already been increased through
length. In addition, although working memory capacity predicts performance on multiple-choice and
recall comprehension items, it does not interact with length or density; this suggests that the effects of
length and density cannot be explained through increased working memory load alone.

---

⁴ The type/token ratio is based on the number of unique words (types) in the passage relative to the total number of
words (tokens).
# TABLE OF CONTENTS

Factors affecting second language listening comprehension .................................................. 5

- Passage length ..................................................................................................................... 5
- Information density ........................................................................................................... 6
- Working memory ................................................................................................................. 7
- Summary ............................................................................................................................... 8

Study 1: Effect of passage length on non-native listening comprehension ............................. 8

Materials ...................................................................................................................................... 8

- Passages ............................................................................................................................... 8
- Length manipulations ......................................................................................................... 9
- Density manipulations ......................................................................................................... 9
- Passage recording ............................................................................................................... 11
- Speech rate manipulations ................................................................................................. 11
- Passages used in Study 1 ..................................................................................................... 12

L2 listening comprehension measures ................................................................................... 12

- Comprehension questions ................................................................................................. 12
- Recall items ......................................................................................................................... 13
- Comprehension measures used in Study 1 ......................................................................... 13

Versant™ Pro Speaking test .................................................................................................. 13

Language Experience and Proficiency Questionnaire (LEAP-Q) ............................................. 13

Participants ............................................................................................................................ 14

Procedure ................................................................................................................................ 15

Data analysis .......................................................................................................................... 15

- Motivation for using Bayesian methods ............................................................................. 16
- Prior distributions ............................................................................................................... 16

Results ..................................................................................................................................... 16

- Versant™ Pro Speaking test ............................................................................................... 17
- Demographic information .................................................................................................. 17
- Multiple choice ..................................................................................................................... 17
- Recall .................................................................................................................................... 19

Discussion ............................................................................................................................... 21

Study 2a: Effects of passage length, information density, and working memory on native listening comprehension ............................................................ 22

Method ..................................................................................................................................... 22

- Participants ........................................................................................................................ 22
- Materials .............................................................................................................................. 22
- Procedure ............................................................................................................................. 24

Data analysis .......................................................................................................................... 24

- Results .................................................................................................................................. 25
- Demographic information ................................................................................................. 25
- Working memory ................................................................................................................ 25
- Multiple choice ................................................................................................................... 25
- Free recall ............................................................................................................................. 28

Discussion ............................................................................................................................... 32
Study 2b: Effects of passage length, information density, and working memory on non-native listening comprehension ................................................................. 32

Method ........................................................................................................................................ 32

Participants ............................................................................................................................... 32

Materials ................................................................................................................................... 33

Procedure ................................................................................................................................. 33

Data analysis ............................................................................................................................ 34

Results ............................................................................................................................................. 35

Versant™ Pro Speaking test .................................................................................................... 35

Demographic information .......................................................................................................... 35

Working memory..................................................................................................................... 35

Multiple choice .......................................................................................................................... 36

Free recall ................................................................................................................................ 39

Discussion ................................................................................................................................. 42

Conclusions ................................................................................................................................. 42

References................................................................................................................................... 44

Appendix A: Sample passage (medium length, low type/token ratio) ........................................ 47

Appendix B: Length of passages (number of syllables) .............................................................. 48

Appendix C: Density of passages (type/token ratio calculated using MATTR) ............................. 49

Appendix D: Sample multiple-choice comprehension questions ........................................... 50

Appendix E: Sample recall items ................................................................................................. 51

Appendix F: Modified version of the LEAP-Q .......................................................................... 52

Appendix G: ILR self-assessment of foreign language listening proficiency ............................. 54

Appendix H: Details of Bayesian statistical analysis .................................................................. 53

Specification of prior distributions ............................................................................................. 57

MCMC details ............................................................................................................................... 58

Model interpretation ................................................................................................................... 58

Imputation procedure ................................................................................................................ 58

Summary .................................................................................................................................... 57

APPENDIX I: Demographic descriptors for all studies ................................................................. 58

Study 1 ......................................................................................................................................... 58

Study 2a ...................................................................................................................................... 58

Study 2b ...................................................................................................................................... 59
FACTORS AFFECTING SECOND LANGUAGE LISTENING COMPREHENSION

Listening in a second language (L2) is challenging for a number of reasons. Listening involves real-time processing, generally without the option of referring back to earlier sections of the passage (Buck, 2001; Cutler, Dahan, & van Donselaar, 1997). While most reading situations allow the reader to have complete control of the rate at which text is received, listeners have much less control over the speed of delivery (Osada, 2004). In addition, in comprehending spoken language, word boundaries must be inferred from a variety of lexical and phonological cues (e.g., Cutler et al., 1997). While these demands are fundamental to the act of listening, other factors that may affect comprehension difficulty vary across passages, such as speech rate and the presence of infrequent vocabulary. It is important to systematically investigate how these variable factors impact listening comprehension to determine how difficult a given passage will be for an L2 learner to understand.

Most L2 comprehension research has focused on reading rather than listening, possibly because the process of reading is more easily observed and manipulated (Osada, 2004). Thus, there remains a great deal that is not known about how qualities of a passage affect listening comprehension for L2 learners. The lack of research in this area is particularly problematic for language instructors, who select and create classroom listening materials, and for language test developers, who must determine the difficulty of listening materials. Below, we briefly review the available research on how passage length and information density impact L2 listening comprehension. In addition, we discuss research examining the role of working memory capacity in L2 listening comprehension and discuss how individual differences in working memory capacity may determine the impact of length and density.

Passage length

One factor of concern in L2 listening comprehension is the extent to which listeners can cope with the amount of information that is presented for processing (Alderson et al., 2006; Bejar, Douglas, Jamieson, Nissan, & Turner, 2000; Carroll, 1977, as cited in Dunkel, 1991, p. 440; Dunkel, 1991; Rost, 2006). As mentioned above, listening comprehension is unlike reading because it occurs in real time. Generally speaking, listeners do not have the option of going back to something they failed to comprehend. Instead, the result of such a failure will be an inability to attend to new information as listeners invest additional time in attempting to understand what they missed, or an inability to comprehend later information because it relies on the understanding of earlier missed information (Goh, 2000; O’Malley, Chamot, & Küpper, 1989). The more information presented by a passage, the more opportunities are present for a derailment of comprehension and the more cases of interdependency between pieces of information a listener is likely to encounter. Longer passages may also be more likely to overwhelm listeners’ working memory storage capacity by providing a larger amount of information for the listener to retain and integrate with incoming information (Henning, 1991).

Despite these compelling arguments for an effect of passage length, research studies examining the effects of passage length on L2 listening comprehension suggest that length is not a strong factor in predicting difficulty (e.g., Nissan, DeVincenzi, & Tang, 1996). Studies that do find a relationship find one that is weak (Henning, 1991) or qualified by interactions (Carrell, Dunkel, & Mollaun, 2002). One reason for the lack of consistent effects of length on listening comprehension may be that the range of compared lengths varies greatly across studies (e.g., 2.5 versus 5 minutes in Carrell et al., 2002; 10 words versus 30 words in Henning, 1991). In addition, many studies investigating length did not manipulate the length of listening materials independently of other factors that may impact comprehension difficulty (e.g., Moyer, 2006), making it impossible to determine how much of the observed effect (if there is an effect) was due to differences in passage length.

While the lack of consistent results may be due to problems of study design and differences in the range of lengths examined across studies, this pattern may also be due to passage length’s relationship to other variables like redundancy (when information is presented more than once through repetition, elaboration, or other methods). Research findings indicate that additional information that is redundant
helps rather than hinders L2 listening comprehension. Gainer (1997) found that dialogue passages where key information presented by the first speaker was echoed by the second speaker (e.g., Speaker 1: He was born in 1955. Speaker 2: Born in 1955), yielded superior comprehension for both higher- and lower-proficiency listeners, compared to a shorter, unmodified version of the same passage. Chiang and Dunkel (1992) explored the effects of redundancy presented in the form of paraphrasing (The food of the Pennsylvania Dutch Country is very hearty and delicious. Hearty and delicious food is nourishing and tasty, p. 354) and found that this redundancy improved comprehension for higher-proficiency listeners, even though the passage containing redundant information was longer.

It is important to note, however, that not all redundant information will be equally beneficial to comprehension or helpful for listeners of all levels of proficiency. Chaudron (1983) examined the effects of five types of topic restatements, with differing syntactic complexity and psychological salience, on L2 listening comprehension. Participants showed the worst performance on comprehension items when redundancy was presented in its least salient form (synonyms: They are selling beer at the picnic. The brew tastes terrific, p. 441), but higher-proficiency participants understood more than did medium- or lower-proficiency participants when hearing more complex types of redundancy (such as nouns repeated in rhetorical questions [What about the beer? It tastes terrific, p. 443]). Lower- and medium-proficiency listeners benefitted most from exactly repeated nouns (What about the beer? The beer tastes terrific, p. 441). Gainer’s findings are consistent with Chaudron’s (1983) finding that redundancy in the form of immediate, exact repetition (high saliency redundancy) is beneficial for both lower- and higher-proficiency listeners. In turn, Chiang and Dunkel (1992) found that paraphrase, a more complex form of redundancy (per Chaudron, 1983), benefitted comprehension for higher but not lower-proficiency listeners, consistent with Chaudron’s (1983) findings. Thus, when considering whether the length of a passage is due to redundant information or unique information, it is also important to consider the type of redundancy to predict how difficult the passage will be for L2 listeners to comprehend.

**Information density**

One rationale behind examining passage length as a factor in L2 listening comprehension is the belief that a greater processing load is introduced by longer passages (Carroll, 1977, as cited in Dunkel, 1991, p. 440; Dunkel, 1991, p. 440; Henning, 1991; Rost, 2006). If this is the explanation for passage length’s effect, how much information a passage presents, rather than just word or syllable count, may be a more predictive factor for comprehension difficulty than is overall passage length. Information has been defined several ways in the literature: content words (e.g., a noun, verb, adjective, or adverb, Nissan et al., 1996); words with independent meaning (e.g., mother vs. a, Gilmore, 2004); and propositions (the smallest unit of knowledge that can stand alone as a separate true-false statement, Dunkel, Henning, & Chaudron, 1993). Most measures of information density involve dividing the number of pieces of information in a passage by the total number of words in the passage or the duration of the passage (i.e., capturing how packed a passage is with information, given its length). Sometimes measures of density only include those pieces of information that have not been previously given in the passage in the numerator (i.e., they control for redundancy, Aiken, Thomas, & Shennum, 1975); these measures directly capture the density of the unique information in a passage. Like passage length, greater information density is believed to make higher cognitive demands of L2 listeners, which in turn increases the effort involved in listening comprehension (Gilmore, 2004).

Buck and Tatsuoka (1998) found that the proportion of content words to all words surrounding the item-necessary information was a significant predictor of item difficulty. Rupp, Garcia, and Jamieson (2001) explored how information density in a passage or text affected the difficulty of L2 listening and reading comprehension items and found information density to be an important variable for predicting item difficulty. They examined the relationship between type/token ratio (TTR), a measure of information density, and comprehension difficulty for a reading task and a listening task. TTR emerged as a significant predictor of item difficulty, with test items for passages with higher information density being more difficult for the listener.
TTR is a measure of lexical diversity that represents the proportion of *types* (unique words) to *tokens* (all words) in a passage. However, TTR can also be construed as a measure of information density that accounts for redundancy, as the count of types is dominated by content words over function words (e.g., *shirt* is less likely to be repeated in a particular passage than is *her*, *the*, or *however*). Further, some manners of defining *unique* for the purpose of counting types can lead to a TTR that represents more the density of novel information in a passage rather than simply the density of unique words in a passage. For instance, Richards (1987) presented a list of rules for defining types that excludes inflected forms of already presented words (e.g., *running* is not a new type if *run* is already present; *be* is not a new type if *was* is present). This manner of defining type results in synonyms for previously presented words (e.g., *run* and *race*) but not inflected forms of previously presented words (e.g., *run* and *running*) increasing the diversity of a passage. Given that non-native speakers are less sensitive to more complex forms of redundancy (e.g., Chaudron, 1983), it is likely that non-native speakers may not recognize that synonyms relay redundant information. Thus, TTR (with type defined in this more exclusive way) may represent the density of novel information in a passage for non-native speakers (more information about the way type was defined in the current studies is provided in the *Density Manipulations* section).

None of the above studies directly manipulated the level of information density in passages while holding other factors constant; rather, they examined passages where this factor freely varied along with a number of other factors, such as topic and vocabulary. However, relevant research suggests that information density does impact listening comprehension. As with length, an investigation of how information density impacts L2 listening comprehension when other relevant factors are controlled is necessary to clarify the importance of this factor.

*Working memory*

Understanding a foreign language requires several general cognitive abilities. One such ability is working memory, typically defined as the capacity to attend to, temporarily store, and process incoming information. Listeners with a greater working memory capacity understand more of what they hear when they are listening to their non-native language than listeners with a lower working memory capacity (Harrington & Sawyer, 1992). Miyake and Friedman (1998, p. 348) described a study performed by Miyake, Friedman, and Osaka (1998) exploring the causal relations between working memory and L2 listening comprehension. The results revealed an impact of working memory on syntactic processing: participants with higher working memory capacity were able to make better use of syntactic information when comprehending the L2 and demonstrated a level of sensitivity to particular syntactic cues that was near native-listener levels. Further evidence for the role of working memory on syntactic processing was reported by McDonald (2006), who found that working memory correlated significantly with the accuracy of grammaticality judgments of spoken L2 sentences. In sum, the available research evidence suggests that individuals’ working memory capacity affects L2 listening comprehension.

The factors discussed above (rate, length, density, and redundancy) are likely to impact the L2 listener’s available working memory. In general, the more information held in working memory during a task, the greater the demands on working memory (Baddeley & Hitch, 1974). Passages with longer length may overwhelm working memory for L2 listeners, because it presents a large amount of information to be stored and manipulated. Because of the challenges for working memory already presented by L2 listening (e.g., retrieval of the meaning of L2 words), the effect of length may be more apparent for L2 listeners than it is for L1 listeners. Similarly, passages with greater information density will likely increase working memory demands during listening comprehension by presenting more information for processing in the same amount of time. However, in considering how passages of a longer length or greater density tax working memory, it is important to consider whether the information is unique or redundant. Redundancy, because it reinforces previously given information, may decrease working memory demands, particularly if the form of redundancy has low complexity and high salience (i.e., it is easily recognized as redundant) or if the listener has higher L2 listening proficiency. Furthermore, the effect of a passage-based factor like length or density on L2 listening comprehension is likely to depend on the working
memory capacity of the listener; while a listener with higher working memory capacity may experience little change in comprehension difficulty as information density increases, a listener with lower working memory capacity may be greatly affected by increases in density.

Summary

Although relevant research in the L2 listening comprehension literature suggests that length is not a strong determiner of comprehension difficulty, existing studies have not explored the impact of length when other factors were held constant across passages. Similarly, although previous studies have found some support for the role of information density in L2 listening comprehension difficulty, these studies have failed to manipulate density directly while controlling for other aspects of the listening passages. To address the gaps in the literature and investigate the effects of length and density in a more controlled design, we conducted three studies. In Study 1, we studied the effects of passage length on L2 listening comprehension while holding constant other factors, like topic, number of speakers in the passage, information density, and speech rate. In Study 2a, we examined the impact of information density, defined as the type/token ratio (TTR) of the passage, on listening comprehension for short, medium, and long passages with native speakers. In Study 2b, we replicated the design of Study 2a with non-native speakers. Further, in both Studies 2a and 2b, we examined individual differences in working memory capacity and how the effects of passage length and information density changed as listeners’ capacity increases.

We selected TTR as our operationalization of information density for two reasons. First, although TTR is generally depicted as a measure of lexical diversity rather than density, we defined type in line with Richards (1987) so as to capture more than simply the exact replication of previously given words. Second, we wished to independently manipulate density and length of the passage while maintaining the basic information conveyed by the passage. It is nearly impossible to manipulate these two factors independently when using other measures of density without also altering the ideas conveyed in the passage. Replacing a content word with a function word (i.e., defining information as in Nissan, et al., 1996) or removing a proposition (i.e., defining information as in Kintsch, 1974) invariably changes the meaning of the passage or reduces its length.

STUDY 1: EFFECT OF PASSAGE LENGTH ON NON-NATIVE LISTENING COMPREHENSION

Previous investigations of the impact of passage length on L2 listening comprehension have produced mixed results. However, these studies have not fully controlled for other factors likely to impact listening comprehension, such as passage topic, speaker, or speech rate. The purpose of Study 1 was to examine the impact of length, in syllables, on listening comprehension measures when other factors are held constant.

MATERIALS

Passages

All three studies reported here were based on the same set of listening materials. We created these materials to systematically and independently manipulate length (in syllables) and density (TTR) while maintaining passage topic, speech rate, and speaker(s) constant. While this study uses only the passages with a medium density and varying in length, we will describe the process by which we developed the full set of materials.
The original listening passages were selected from the set (audio and transcripts) presented in the National Foreign Language Center’s (NFLC’s) Introduction to Passage Rating course (2005). NFLC documentation leveled four of these passages at ILR Level 2 and six at ILR Level 3 proficiency, according to the NFLC.¹ Researchers selected two additional passages from WTOP (a local radio station in Washington, DC) news broadcasts and leveled these internally as ILR 2, with that level confirmed by a DLIFLC passage rating expert. All of the passages are similar to the types of passages used on the DLPT, which is the official test of record for foreign language proficiency for military and government employees. The original passages were genuine samples of spoken English recordings, in that they were created for native speakers by native speakers (i.e., they were not created for the purpose of educating non-native speakers in English or with the explicit intent of being intelligible to non-native speakers) and were not created for testing purposes. A sample passage can be found in Appendix A.

The manipulations of length and density described in this section involved working with the transcripts of the NFLC passages. When professional voice talent recorded the passages, they listened to the audio recordings of the original passages and emulated, as closely as possible, the speaking style of the original speakers.

**Length manipulations**

Researchers created three different length versions of each of the 12 base passages: short = 213–243 syllables; medium = 328–398 syllables; and long = 699–756 syllables. The original passages differed in length; in cases where the base passage did not contain enough syllables for the long-length version (i.e., when its length was below the targeted range of ~700–750 syllables), additional content on the same topic and in the same style was added to the passages with the intent of avoiding information redundant with previously presented information. To obtain shorter length versions from a longer base passage, segments of the passage were deleted with an effort to maintain the coherence of the passage. Deleted and added information was distributed across the entire passage and attention was paid to ensuring that each length version conveyed the same basic information to the listener. The research team evaluated the length manipulations to ensure that all versions were coherent and not excessively redundant.

Appendix B shows the number of syllables for each version of the individual passages, based on counts generated from www.syllablecount.com.

**Density manipulations**

The type/token ratio is based on the number of unique words (types) in the passage relative to the total number of words (tokens). TTR controls for exact repetition by counting each occurrence of a word as a type only once. However, as discussed above, this measure can go beyond simply representing the redundancy of a passage by treating words from the same word families as instances of the same type (e.g., the appearance of the words dog and dogs would be counted as two tokens of one type), as in Rupp et al. (2001), or by treating all transformations of a root word as tokens of a single type (e.g., is, are, been, etc. as tokens of be, as in Richards, 1987). There are several methods of defining type, from treating each new lexical item that is not perfectly identical to a previous item as a new type (e.g., dog and dogs would be counted as different types) to more stringent definitions that take into account inflection and other transformations of the root, counting versions of the root as tokens of the same type (e.g., help, helps, helping will count as three tokens of a single type). The definition of type used in the current passage manipulations largely adopted the set of rules presented by Richards (1987, p. 204):

---

¹ More information about the NFLC passage rating course may be found at [http://www.nflc.org/projects/language/pr#.Ud2MLZyDn-A](http://www.nflc.org/projects/language/pr#.Ud2MLZyDn-A). Further information about the abilities involved at each ILR level can be found at [http://www.govtilr.org/skills/ILRscale3.htm](http://www.govtilr.org/skills/ILRscale3.htm).
• Inflected and non-inflected forms (e.g., helping and help) of the same stem count as a single type.
• Contractions of subject and predicate (e.g., I’m, she’s) are treated as two words. Full and contracted forms (e.g., am as it forms a part of both I am and I’m) are treated as a single type.
• Contractions of the verb and negative particle (e.g., can’t) count as one token. These verb forms count as separate types from the corresponding affirmative forms.
• Possessives are different types than names without the possessive ‘s. Thus, Sarah and Sarah’s would count as different types.
• Hyphenated words count as one word.

In contrast with Richards (1987), the definition of type we used included interjections (e.g., oh) in order to be consistent with the decision to include hesitations (e.g., ah, um) and false starts (e.g., Well I-…Well, that’s…), elements that Richards did not address. We believe these non-lexical items should be included in any calculation of TTR because they may not be familiar to L2 listeners, and thus may pose the same comprehension challenge for these listeners as do lexical items. Further, non-hyphenated compound nouns (e.g., United Nations) were not treated as a single type, unlike in Richards (1987), in part because L2 learners may not have lexicalized these compounds. Finally, for ease of calculation, the type definition adopted for the passage manipulations did not distinguish between verb forms in their auxiliary or main verb roles (e.g., do).

Researchers created two TTR levels of each of the three length versions of the twelve passages: medium density (0.59–0.65) and high density (0.69–0.74).

• To decrease the TTR, unique words were replaced with tokens of previously given types (e.g., assisting was replaced with helping if the verb help was already present in the passage).
• To increase TTR, tokens of previously given types were replaced with unique types (e.g., helping was replaced with assisting if the verb help was already present in the passage).

TTR can be manipulated while maintaining the syllable count in the passage, as shown below:

Base passage

The United Nations refugee agency cares for nearly 20 million people who were forced to flee their homes because of war. More than a third of them are between the ages of 12 and 24. The agency says young refugees must be given every opportunity possible to develop their potential. They must be given the help and protection they deserve. More than 70 countries are marking World Refugee Day with special events. Some are holding rock or classical music concerts. A number of local and worldwide celebrities are lending their fame to publicize the plight of refugees—young and old.

Tokens replaced with types (increased TTR)—changes are underlined

The United Nations refugee agency provides support for nearly 20 million people who were forced to flee their homes because of war. A large number of them are between the ages of 12 and 24 years old. The group says young refugees need to receive every opportunity possible to develop their potential. They must be given the help and protection they deserve. Over 70 different countries are marking World Refugee Day with special events. Some are holding rock or classical music concerts. Many local and worldwide celebrities are lending their fame to publicize the plight of refugees—of all ages.

Types replaced with tokens (decreased TTR)
The United Nations refugee agency cares for some 20 million people who had to flee their nations because of war. More than a third of the refugees are between the ages of 12 and 24. The agency says young refugees must be given every opportunity possible to develop their potential. Younger people must be given the help and protection they deserve. More than 70 nations are marking World Refugee Day with special events. Some are having rock or classical music events. Some local and worldwide celebrities are giving their help to publicize the plight of refugees—young and old.

The base passage has 100 words, 160 syllables, and a TTR of 0.68. To increase TTR (the second version), 20 words with 25 syllables were replaced with 17 words with 25 syllables, introducing eight new types and bringing the type/token ratio to 0.78. In the third version, TTR was decreased by replacing 15 words that had 23 syllables with 14 words that had 23 syllables. This removed 10 types and brought the TTR to 0.59. Note that because these passages are very short (~100 words, 160 syllables), it is more difficult to decrease the TTR and the effects of doing so are more obvious because there are simply fewer types to choose from when replacing types with tokens. Further, TTR calculated across an entire passage has a strong negative correlation with passage word count (see, e.g., Covington & McFall, 2010; Malvern & Richards, 2002), resulting in a relatively high TTR for the base passage in the example above.

Appendix C shows the densities (TTRs) for each individual passage. These TTRs were calculated in two steps. First, the original passage transcript was submitted to a machine stemmer and part-of-speech tagger called Machinese 4.9.1 at http://www.connexor.ed/technology/machinese/demo/tagger/index.html. This program automatically converts words to their uninflected forms. Researchers manually checked all of the Machinese output for accuracy and to remove extraneous punctuation. Next, researchers used the Moving Average Type/Token Ratio calculator (MATTR; Covington & McFall, 2010) to measure TTR. MATTR calculates TTR iteratively for passage sections (windows) of a fixed token length, moving one word forward in the passage for each successive window; final TTR level is the average across all windows. This method of measuring TTR greatly reduced the relationship between passage length and TTR (Covington & McFall, 2010). The length and density manipulations resulted in the creation of six different versions of each of the 12 passages—two densities (medium and high) for each of the three lengths.

**Passage recording**

We employed professional voice actors to record each of the six versions of the set of twelve passages (72 passages total). The same voice actor played the same role in each version of each passage, to control for potential effects of speaker across length and density versions. We provided recordings of the original NFLC passages to voice actors and instructed them to emulate the original speaker as much as possible. Passages were re-recorded with voice actors as needed to attain an acceptable level of comprehensibility and the appearance of authenticity. Test developers from DLIFLC approved the final versions of these passages.

**Speech rate manipulations**

We first needed to identify a speech rate for passages for use in the current studies which would yield acceptable levels of comprehension for non-native speakers of English (i.e., not too fast to greatly impede comprehension) yet would sound natural to native speakers (i.e., not too slow as to sound strange or inauthentic). Previous research has reported a wide range of acceptable or ideal speech rates for non-native speakers, though these studies did not investigate whether the “best” speech rate was one that sounded authentic to native listeners.

Two pilot studies were used to identify the speech rate to be used in the present studies. First, 15 native speakers of English were asked to rate the comprehensibility, naturalness, and perceived speech rate of the twelve listening passages when presented at five different speech rates (3.0–5.0
Results indicated that participants were sensitive to the speech rate manipulations: ratings of how fast the passage sounded increased linearly with speech rate. Ratings of ease of comprehensibility and naturalness indicated that the pilot participants found both the slower and faster passages harder to understand. Based on these results, we selected the speech rates of 3.5 syllables per second, 4.0 syllables per second, and 4.5 syllables per second for use in a second pilot study involving non-native speakers. The medium-length, medium-density version of the twelve passages described above were presented to 34 non-native English speakers from the same L1 background (Arabic) at 3.5, 4.0, or 4.5 syllables/second. Pilot participants performed best on comprehension items for passages presented at 4.0 syllables/second, so we selected this speech rate for use in Studies 1, 2a, and 2b. This speech rate is much faster than those identified as ideal for non-native speakers in the studies of British English (between 1.93 and 2.85 syllables per second; Griffiths, 1990) and Hebrew (3 syllables per second; Rosenhouse, Haik, &Kishon-Rabin, 2006), but it is slightly slower those that observed by Derwing and Munro (2001) when they tested comprehension of American English by non-native speakers (4.9 syllables per second).

The relative speech rates of the recordings of each length and density version of the 12 passages (6 versions of each passage, total) were adjusted to 4.0 syllables per second using the Pitch Synchronous Overlap and Add (PSOLA) algorithm (Moulines & Charpentier, 1990). This resulted in short passages of 53–61 seconds, medium passages of 82–100 seconds, and long passages of 175–189 seconds.

Passages used in Study 1

Study 1 assessed the impact of length on L2 listening comprehension when other factors (e.g., density) were held constant. Study 1 presented participants with the short, medium, and long versions of each passage having the medium level of information density (TTR).

L2 listening comprehension measures

All three studies presented participants with the same set of multiple-choice comprehension items (2 per passage). In addition, participants in each study were asked to recall a small number of words (4–6) for each passage; due to differences in the actual verbiage used between length and density versions of the passage, these items differed slightly across passage versions.

Comprehension questions

The L2 Listening Comprehension team worked closely with DLIFLC to ensure that the comprehension test items resembled those that appear on the DLPT5.

Researchers developed the comprehension items such that they would be appropriate for all length and density versions of the passage. Two items were developed per passage. Using the “DLPT5 Lower Range Multiple-Choice Test Specifications” document as a guide, an effort was made during item development to match the item type distribution (e.g., the number of comprehension items that target global vs. non-global information in the passage) and skills (e.g., the listeners’ ability to understand the main ideas, major details, and implications of the passages) recommended for Level 2 and Level 3 passages that assess listening proficiency in the DLPT item development specifications. The multiple-choice comprehension items were reviewed by a CASL researcher with extensive experience in second language test development (Steven Ross).

Researchers conducted two rounds of pilot testing with a small number of native speakers to determine if any of the items could be keyed without listening to the passages. Items were eliminated or revised based on the results. For Study 1, the finalized English items were translated into Modern Standard Arabic (MSA), reviewed by two highly proficient Arabic readers, and piloted with five native Arabic speakers to ensure that the translated items could not be keyed without listening to the passages. For Study 2b, the finalized English items were translated into Spanish, reviewed by two highly proficient
Spanish readers, and piloted with five native Spanish speakers. See Appendix D for examples of multiple-choice comprehension questions.

Recall items

Tasks involving recall of specific words presented in listening materials have been argued to be somewhat unnatural (Hummel, 1993). Nonetheless, recall tasks have been used in several previous studies to measure the listening comprehension of both native and non-native listeners (e.g., Brett, 1997; Cervantes & Gainer, 1992; Conrad, 1985; 1989). Further, ability to recall the exact words used in a listening passage is facilitated by knowing that a recall test is forthcoming (Gurevich, Johnson, & Goldberg, 2010).

Recall items displayed excerpts from the passage and required participants to fill in missing words (indicated by a blank space in the phrase or sentence) using exactly the same words heard in the passage. Items appeared in the same order as presented in the passage.

Because the length and TTR manipulations involved removing, adding, or replacing words in the passages, it was impossible to create one identical set of recall items applicable for all six versions of the passage. Instead, researchers developed sets of four to five recall items (contexts plus blanks for the cued words) for each density level of a given passage. The root word (e.g., kick with one version cuing kicking and another cuing kicked) cued for each item was identical across the three levels of length for each passage, but the context provided for the cued word varied somewhat because of differences in the actual words spoken in the different length passages. Whenever possible, the same root word was cued for all three density versions of a passage. Cued words were content words selected to be more likely to be known by participants (i.e., not low frequency words or proper names).

Recall items were piloted with five native speakers of English to determine if the cued word could be guessed without having listened to the passage, and revisions were made as necessary. All recall items were in English for all three studies. An example of the six versions of a finalized recall item is provided in Appendix E.

Comprehension measures used in Study 1

Study 1 presented participants with the short, medium, and long versions of each passage having the medium level of TTR. Accordingly, Study 1 used the recall items corresponding to these versions of the passages and the multiple-choice items, which were common to all six (3 length x 2 TTR) versions of each passage, to measure comprehension.

Versant™ Pro Speaking test

Researchers included the Versant™ Pro Speaking test in the experiment materials to establish participants’ English listening proficiency. All instructions for the Versant™ Pro test were translated into Modern Standard Arabic (MSA) for Study 1 to ensure acceptable comprehension of the tasks.6

The Versant™ Pro test took approximately 25 minutes to complete and involved eight tasks: Read Aloud, Repeats, Short Answer Questions, Sentence Builds, Story Retelling, Response Selection, Conversations, and Passage Comprehension (http://www.versanttest.com/products/proSpeaking.jsp; Pearson, 2010). The latter three tasks contributed to the listening comprehension subscore. In Response Selection, participants use a multiple-choice format to select the most appropriate response to a spoken utterance. In Conversations, participants listen to a conversation followed by a question and then answer the question with a few words. In Passage Comprehension, participants listen to a passage followed by

6 An English version of the Versant™ Pro test instructions was on hand during all experiment sessions if the participants preferred to read the instructions in English.
three questions and then answer each question with a few words. Although the last two sections require participants to speak their responses, no part of the test involves a dialogue between speakers, and thus the listening comprehension assessment is best described as assessing non-participatory listening. Indeed, every task in the Versant™ Pro is machine-scored and administered over the telephone. The listening comprehension subscore was used as an objective measure of English listening comprehension level for the participants in Studies 1 and 2b. Participants in Study 2a were all native speakers of English who did not take the Versant™ Pro Speaking test.

All participants in Studies 1 and 2b received instructions on how to acquire their Versant™ Pro Speaking test score report at the end of their experiment session. The report included an overall score as well as scores for Listening Comprehension, Sentence Mastery, Vocabulary, Fluency, and Pronunciation.7

Language Experience and Proficiency Questionnaire (LEAP-Q)

We included a modified version of the Language Experience and Proficiency Questionnaire (LEAP-Q; Marian, Blumenfeld, & Kaushanskaya, 2007), translated into MSA for Study 1, and into Spanish for Study 2b, to obtain further information about the language background of the participants. This questionnaire includes items about age of acquisition for any foreign languages, number of years of formal education, and current use activities with foreign languages. The English items of the LEAP-Q used in the current studies are presented in Appendix F.

Participants

Fifty-two native Arabic speakers (46 male, 6 female) learning English participated in Study 1. The majority (31) of these participants were native speakers of Arabic studying English at the Defense Language Institute’s English Language Center (DLIELC) in San Antonio, Texas. For these participants, instructors of the DLIELC’s English courses distributed the researchers’ contact information to those students whom they believed to have the appropriate level of English listening proficiency for the study (i.e., ILR 2/2+).8 One additional participant was recruited in San Antonio from outside the DLIELC community. The remaining participants (20) were recruited in the College Park, Maryland area through fliers, word of mouth, and contact with organizations likely to include a large proportion of native Arabic speakers. Initial recruiting efforts did not employ any formal English language proficiency screener, stipulating only that participants have an intermediate level of English proficiency; when recruiting efforts shifted focus to a professional marketing firm, a phone screener was used to eliminate participants who had a level of English proficiency too high for the current studies. This screener involved a set of self-assessment items adapted from the ILR Self-Assessment of Foreign Language Listening Proficiency. The ILR self-assessment consists of Yes/No can-do statements for different language levels (the screening protocol in the current study used statements for Levels 1–3). Responding “Yes” to 9–15 of the 17 self-assessment items qualified an individual for participation in the study. The set of self-assessment items is included in Appendix G.

All participants were required to take the Versant™ Pro Speaking test prior to the study. Listening comprehension subscores on the Versant™ Pro Speaking test indicated that the English listening proficiency of six of the participants was approximately ILR Level 1. In addition, the proficiency of three participants was approximately ILR Level 3. Forty-three of the participants were at approximately ILR

---

7 Researchers and participants accessed all Versant™ score information via the Versant™ website: http://versanttest.com/.
8 Level 2 in listening is described as Limited Working Proficiency (“Interagency Language Roundtable Language Skill Level Descriptions: Listening”, 1985).
Level 2 in their English listening proficiency according their Versant™ Pro Speaking test listening comprehension subscore. 

**Procedure**

As noted above, participants took the Versant™ Pro Speaking test prior to the experiment session. Consistent with the Versant™ test guidelines, participants were instructed to take the test using a landline telephone.

Upon arriving at the test site, participants provided informed consent in writing and then completed all experiment materials using an individual computer and headphones.

All task instructions were presented in MSA. First, participants provided their Versant™ Speaking Pro test code in order to link their listening comprehension subscore to their responses. Participants next completed the LEAP-Q and then the portion of the experiment involving listening passages, multiple-choice comprehension questions, and recall items. When administering this portion of Study 1, researchers adhered to the procedure outlined in the Defense Language Proficiency Testing System 5 Framework (2010) as closely as possible. Item layout and response procedures were described in detail in the instructions, and participants responded to a practice passage to familiarize themselves with the passage play button and the format of the multiple-choice and recall items.

For each passage, participants first read an orientation statement in MSA and saw the multiple-choice comprehension questions (presented in MSA), which could be answered at any time. The play button for the passage appeared under the multiple-choice questions, and participants could begin play of the passage whenever they wished. Each passage played twice without pausing, with a tone sounding between repetitions. Once participants finished answering the comprehension questions, they could press the “Next” button to proceed to a screen with the set of recall items for that passage. The passage audio was not available on the recall items screen. Participants typed their answers to the recall items in English and could answer these items in any order. Participants were allowed to take a break halfway through the test.

At the end of the study, participants were thanked, debriefed, and provided with their Versant™ Speaking Pro test scores.

**Data analysis**

Throughout this report, the dependent variable of interest was accuracy, coded as a dichotomous variable (i.e., 1 = correct, 0 = incorrect). Because of the dichotomous nature of the outcome measure, we analyzed the accuracy data via logistic regression. In these models, performance was modeled in terms of the estimated log-odds of making a correct response and the effect of a predictor indicates the corresponding change in log-odds (Gelman & Hill, 2007). Throughout all analyses, we included random intercepts that varied by participant and passage; all experimental factors and covariates were included as

---

9 We defined ILR level 2/2+ as having Versant™ listening comprehension test scores between 41 and 68. We selected this range based on several sources. Bernstein & De Jong (2001) found that the Versant™ test scores predict proficiency levels on the Common European Framework of Reference (CEFR) scale of the Oral Interaction Skills with reasonable accuracy. This information, in combination with a table published by the American University Center of Provence (2012) that provided CEFR equivalencies to ILR levels allowed us to establish that people with Versant™ scores between 47 and 68 were likely to listen at an ILR level of 2. This was corroborated separately when we obtained the Oral Proficiency Interview (OPI) scores of 16 of the students who had volunteered for our study. These participants were all rated by their OPI assessment as being at Level 2 for both listening and speaking; their overall scores on the Versant™ Speaking Test ranged from 42 to 66, while the listening comprehension subtest scores ranged from 41 to 63. Based on these data, we decided that we would select participants who scored between 41 (the lower end of the range for the OPI-tested participants) and 68 (the upper end of the scores observed by Bernstein & De Jong, 2001) on the Versant™ listening comprehension subtest.
fixed effects. In Study 1, the analyses treated length (short, medium, and long) as a categorical variable, and standardized Versant™ scores were included as a simple covariate. Mixed-effects (i.e., hierarchical) logistic regression models were fit to the multiple-choice and recall comprehension data separately.

We implemented the logistic regression models within a Bayesian analytic framework. Details of the model can be found in Appendix H: Details of Bayesian statistical analysis.

Motivation for using Bayesian methods

Bayesian methods offer a number of advantages; the most relevant to this study is the ability to draw inferences regarding an effect being functionally equivalent to zero—i.e., one can affirm the null hypothesis—by estimating the probability that a model parameter is zero or very near zero. This is done by setting a Region of Practical Equivalence, or ROPE, which is a small range of values that are considered to be equivalent in practice to the null value (i.e., a null range). If the 95% highest density interval (HDI; a Bayesian confidence interval) falls entirely within the ROPE, then we can conclude that the parameter is practically equivalent to zero. This is critical to the current study. One of the hypotheses motivating this study was that passage length would not be a significant predictor of listening comprehension. In other words, we hypothesized that the effects of length would be zero, which we can estimate by examining the posterior distribution of the length effect parameters. To define the ROPE for all three studies in this report, we assumed that if a switch between two conditions (or, for working memory, a change of one standard deviation in working memory ability) corresponded to a maximum change of ±5% accuracy or less, this would be considered a minimal change and would be practically equivalent to a null effect. This corresponds to a beta parameter of 0.20 (since .20/4 = .05). Therefore, we set the ROPE at ±0.20 for the length, density and working memory beta parameters across all analyses.

To foreshadow the analysis in Study 2b, the Bayesian analysis also provided a coherent method to address missing data via data imputation. In the final sample for Study 2b, approximately 15% of participants had no Versant™ Listening Comprehension subscore due to a variety of factors (e.g., failure to provide sufficient responses for scoring, poor telephone connection, excessive background noise, speaking too softly). It is well established that simpler methods for addressing missing data—either by dropping any participant with missing data (i.e., case deletion) or selecting a single value to replace each missing data point (e.g., mean replacement, hotdecking)—introduce bias to the resulting model parameters that negatively impacts the validity of the analysis (e.g., Rubin, 1987, 1996). A more valid alternative is to develop an imputation model to predict plausible values for each missing data point that are conditioned on all observed data values, and then the analyst can incorporate these imputed values into the substantive analysis. For details and validation of our imputation modeling approach, see Appendix H.

The Bayesian mixed-effects logistic regression models were fit using the R2WinBUGS package (Sturtz, Ligges, & Gelman, 2005) within the R statistical computing environment (R Development Core Team, 2011). For a detailed description of the Bayesian modeling methods and model interpretation, see Appendix H.

Prior distributions

In Bayesian analyses, the analyst must define prior distributions (or “priors”) for each model parameter. The priors represent the state of knowledge or expectation regarding the model parameters prior to data collection and analysis. More or less information can be integrated into the analysis by setting more or less precise priors. If there is a substantial amount of evidence from previous research regarding the magnitude and direction of an effect, this information can be represented by setting a very precise (i.e., narrow) prior distribution on the effect, indicating that the analyst has a justifiable a priori expectation about the effect. On the other hand, if there is little previous information about the effect, or if the nature of the effect is uncertain, then the analyst can set a wide (or “weakly informative”) prior distribution on the effect. This latter approach simply constrains the model to make unreasonably extreme
parameter values less likely *a priori*, while still letting the data be the primary source of information driving the values observed in the posterior distribution (Kruschke, 2011). For the current analysis, since this is new research, we set weakly informative priors so as to allow the data to drive the results. The detailed specification of the prior distributions and other details on the model fitting procedure are presented in Appendix H.

**Results**

*Versant™ Pro Speaking test*

The distribution of Versant™ listening comprehension subscores for participants in Study 1 is shown in Figure 1 below. Although a small number of participants had scores outside the range corresponding to the targeted 2/2+ ILR level, the majority had scores that fell within the 41–68 range. Prior to analysis, Versant™ scores were centered at a value of 55 (the median value in the range of Versant™ scores corresponding to ILR Level 2 abilities) and standardized.

![Figure 1](image.png)

*Figure 1.* Distribution of Versant™ listening comprehension subscores for participants in Study 1.

**Demographic information**

Participants’ ages ranged from 18 to 62 years, with an average age of 30 and a median age of 28. The majority of participants (75%) reported having at least a college degree, with a sizable proportion of these individuals reporting post-baccalaureate education (~30% of the entire group of participants). No participants reported having less than a high school level of education. The median length of residence in a country where English is spoken was 1 year (range of 0–20 years of residence) and the median time spent in a work or school environment where English is spoken was 4 years (range of 1 month to 21 years). See Tables I-1 and I-2 in Appendix I for more detailed information on demographic descriptors.

**Multiple choice**

Table 1 lists the sample size and by-subject marginal means, standard deviation (*SD*), and minimum and maximum values of the proportion correct for the multiple-choice data. The data show that
participants were able to answer the multiple-choice questions at above chance (.25) levels most of the time.

Table 1. The number of participants for Study 1, along with the descriptive statistics for the percent correct on multiple-choice test items.

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>63.5%</td>
<td>17.3%</td>
<td>20.8%</td>
<td>95.8%</td>
</tr>
</tbody>
</table>

Table 2 provides a summary of the mean percent correct across participants across the three (short, medium, and long) passage length conditions. Differences in length had a minimal impact on performance accuracy when speaking rate and information density were held constant.

Table 2. Mean accuracy (percent correct on the multiple-choice questions) and SD as a function of passage length (short, medium, long).

<table>
<thead>
<tr>
<th>Length</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>64.7%</td>
<td>22.9%</td>
</tr>
<tr>
<td>Medium</td>
<td>62.3%</td>
<td>22.3%</td>
</tr>
<tr>
<td>Long</td>
<td>63.7%</td>
<td>23.3%</td>
</tr>
</tbody>
</table>

The patterns noted above are supported by the logistic regression results. The model coefficients from the analysis are displayed in Figure 2.

Figure 2. Model coefficients for the multiple-choice data in Study 1.
For this analysis, we also included the standardized Versant™ scores as a covariate in the model, in order to control for individual differences in L2 listening comprehension ability. Therefore, for these analyses, the intercept is interpreted as the expected performance in the baseline condition of an individual with a Versant™ score of 55.

In this figure, the points indicate the model parameter estimates and the horizontal lines indicate the 95% highest density intervals (HDIs). A black dashed vertical line marks zero, and the red vertical dashed lines indicate the boundaries of the ROPE of +/-0.20. Starting from the bottom row and moving up, the intercept at the bottom indicates that in the baseline reference condition, a participant with a Versant™ score of 55 would be expected to have a log-odds of success around 0.82, corresponding to around 69% accuracy. The effects of medium and long length were not significantly different from zero, as indicated by the parameter values near zero (for medium and long length, $\beta_s = -0.15$ and $-0.09$, respectively) and the HDIs containing zero. However, L2 listening ability, as measured by the Versant™ score, was positively related to performance, as indicated by the positive parameter value of .28 and the HDI spanning [.07, .54]. These results indicate that while proficiency did predict performance, length did not systematically affect it.

Recall

Accuracy on the recall comprehension test involved evaluating responses for either verbatim correctness or equivalence of meaning to the cued word. Recall item responses were hand scored by team members and were awarded credit (1) or no credit (0). The target word, or a word that was an unambiguous misspelling of the target word, received credit (e.g., againist, aganst, agianst, and aginst were accepted for the target word against), as did multiple-word responses that included the target word or a misspelling of the target word (e.g., Cosby’s comments received full credit for the target word comments), and base forms of inflected target words and inflected forms of base target words received credit (e.g., push was accepted for the target word pushed, and deserts was accepted for the target word desert). In addition, researchers assigned credit to responses that indicated gist accuracy, such as synonyms of the target word, misspellings of synonyms of the target word, and multiple-word responses that included a synonym of the target word received partial credit (e.g., happy, hap, and happy that received partial credit for the target word glad). Hyponyms and hypernyms were not awarded credit (e.g., Metro Center and Rockville did not receive credit for the target word stations, and person did not receive credit for the target word listener). Context was always considered when determining whether responses should receive credit for gist accuracy. When a team member was unsure whether a certain response should receive credit, at least two other team members reviewed the word in question, the target word, and the context. In these instances, responses received credit if at least two out of three team members considered the response to be a synonym of the target word. No credit was awarded to all other words, responses that were not recognizable as words, and blank responses.

Table 3 presents the average recall level for the participants. The maximum proportion correct was .920, which is slightly worse than the maximum performance achieved when answering multiple-choice questions (with a maximum of .958).

<table>
<thead>
<tr>
<th>$N$</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>36.5%</td>
<td>22.7%</td>
<td>0%</td>
<td>92%</td>
</tr>
</tbody>
</table>

Table 4 provides a summary of the mean percent correct across participants for each of the six conditions. Unlike the multiple-choice data, the patterns in the table suggest that passage length had a small but consistent impact on performance, with mean accuracy decreasing for the longest passages.
Table 4. Mean accuracy (percent correct on the multiple-choice questions) and SD as a function of passage length (short, medium, long).

<table>
<thead>
<tr>
<th>Length</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short</td>
<td>37.4%</td>
<td>23.2%</td>
</tr>
<tr>
<td>Medium</td>
<td>37.0%</td>
<td>25.8%</td>
</tr>
<tr>
<td>Long</td>
<td>34.8%</td>
<td>26.6%</td>
</tr>
</tbody>
</table>

Table 4. Mean accuracy (percent correct on the multiple-choice questions) and SD as a function of passage length (short, medium, long).

The patterns noted above are supported by the logistic regression results. The model coefficients from the analysis are displayed in Figure 3.

Figure 3. Model coefficients for the recall data in Study 1.

Starting from the bottom row and moving up, the intercept at the bottom indicates that in the baseline reference condition, a participant with a Versant™ score of 55 would be expected to have a log-odds of success around −0.53, corresponding to around 37% accuracy. The effect of medium length was again not significantly different from zero, as indicated by the parameter value near zero (β = −0.09) and the HDI containing zero. When listening to a long passage, however, the results suggest a small, negative effect (β = −0.19). Although the HDI contains zero, approximately 94% of the posterior distribution of credible values fell below zero, indicating that we can be moderately confident that long passage length negatively impacts performance on recall items. L2 listening ability, as measured by the Versant™, was again positively and strongly related to performance, as indicated by the positive parameter value of .57 and the HDI spanning [.23, .92]. Unlike the results for multiple-choice items, these results suggest that length plays a role in recall accuracy, albeit not a large one.
Discussion

Results for the multiple-choice comprehension items suggest that while English listening proficiency (as measured by the Versant™ Listening Comprehension subtest) strongly predicted accuracy on multiple-choice items, there was no impact of passage length. For recall items, however, accuracy was lower when passages were long compared to short. The difference in length’s role in accuracy for the multiple-choice and recall items may be due to the different demands of these tasks. Participants could read the multiple-choice items prior to listening to the corresponding passage and could answer these items, and change their answers at any time while listening or before moving on to the next screen with the recall items. Under these circumstances, participants could plan to listen for the information necessary to answer the multiple-choice items. In addition, because each passage played twice, participants could have used the second play of the passage to check their previous responses to the multiple-choice items. By contrast, the recall items were unavailable until after participants listened to the passage, so participants were unaware as to what information these items would target until passage play was complete. In addition, the multiple-choice items presented a selection of potential answers to participants, which they could evaluate against their understanding of the passage information, while the recall items were open-ended. Thus, the differential relationship between length and response accuracy in Study 1 is consistent with the previous finding that passage length impacts recall tasks but not recognition tasks (Schultz, Jr. & Johnson, 1982).

Other factors related to the amount of information in a passage, such as the information density, may play a bigger role in listening comprehension difficulty and, thus, in the accuracy of participants’ responses to recall and multiple-choice items. Studies 2a and 2b investigated how density, operationalized as type/token ratio, impacts comprehension. In addition, these studies investigate how density interacts with length and with a participants’ working memory capacity to increase listening comprehension difficulty.

---

10 As mentioned in the Procedure section, we adhered to the procedure outlined in the Defense Language Proficiency Testing System 5 Framework (2010) as closely as possible, which meant that participants saw the multiple-choice comprehension questions before they heard the audio. These questions could be answered at any time. Once the participant had finished with the multiple choice questions, they answered the recall items. These recall items served as an additional comprehension measure that was likely more sensitive than the multiple choice questions.
STUDY 2A: EFFECTS OF PASSAGE LENGTH, INFORMATION DENSITY, AND WORKING MEMORY ON NATIVE LISTENING COMPREHENSION

Building on the design of Study 1, Study 2a investigated the impact of information density on listening comprehension and recall accuracy and the interaction between density and passage length. Study 2a focused on native speakers. There are two reasons for taking this approach. First, inasmuch as there is a paucity of studies exploring the impact of the length of listening materials on comprehension for non-native speakers, there is even less research into how this factor affects performance for native speakers (in fact, we had difficulty locating any empirical studies examining the impact of length of a listening text on native speakers’ comprehension). Second, it was important to establish a baseline to which the likely performance of non-native speakers of English could be compared; if the passages were so long or dense as to greatly impede listening comprehension for native speakers, floor effects might be found for non-native speakers.

Study 2a also explored the extent to which individual differences in working memory influenced recall accuracy and listening comprehension under varying combinations of passage length and information density.

Method

Participants

One-hundred seventy-five participants took part in Study 2a. Recruitment of participants was limited to individuals of 18 -20 years of age. Of the 175 participants, the data from three were excluded due to incomplete working memory data (n = 2) or a lack of consent for us to use his data in the analysis (n = 1). One-hundred seventy-two participants were included in the analysis (116 female, 47 male, 9 unknown). All were native speakers of English studying at the University of Maryland.

Materials

Listening passages

Study 2a was conducted using the full set of passages described in the Methods section of Study 1. Six different combinations of the passages were presented to participants, such that each participant heard one ILR 2 and one ILR 3 passage at each level of length (short, medium, and long) and each level of density (medium and high): 12 passages total. Any given participant heard only one version of a particular passage. The order in which the passages were presented to each participant was pseudo-randomized to avoid order effects.

Comprehension questions

The comprehension items described in the Methods section of Study 1 were used in Study 2a. However, in this case, the original English versions of the multiple-choice items were used instead of the Arabic-translated versions.

Recall test

Study 2a used the full complement of recall items described in the Methods section of Study 1. As described above, since a few words were changed in the passages to match the experimental conditions, this meant that the cued word varied on occasion between density versions, and the context provided for the cued word sometimes differed between length versions according to the exact text presented in the passage.
Working memory capacity assessment

Working memory capacity was assessed using two spatial tasks. One key benefit of these tasks for the current study is that they could be administered to speakers of any language, as they do not rely on the ability to comprehend or produce language beyond reading the task instructions. As the participants in Study 2a were native speakers of English, we provided the original instructions for the task in English. In spatial working memory tasks, participants were asked to simultaneously process and store spatial information. The first spatial working memory task was called Blockspan (Atkins, 2011 as cited in Atkins et al., 2013). In this task, participants saw a 4 x 4 grid of squares and were asked to remember the order and spatial position in which a sequence of yellow blocks appeared on the grid (see Figure 4 below).

Each block within a sequence appeared in yellow for one second, one at a time, in one of the cells on the 4 x 4 grid. Trials were segmented into sets by the appearance of a black square mask that covered the entire grid for one second. After viewing a given trial (displaying a series of locations where a single block flashed in a particular location in the grid), participants were asked to recall the locations in which the squares were flashed in the correct order by clicking the squares in the same order that they appeared.

Participants completed 16 trials of each length (from 2 to 20 blocks in each sequence). The task increased in difficulty by increasing the trial length, and by increasing the number of sets within each trial. For instance, the first trial consisted of one set with 2 stimuli in the trial. For the next trial, there is one set with 3 stimuli, the next trial had one set of 4 stimuli, and the next trial had one set with 5 stimuli. After this, trials were made more difficult by including two sets of 2, 3, 4, or 5 stimuli. Then, trials were made more difficult by including 3 sets of 2, 3, 4, or 5 stimuli. Finally, for the last 4 trials, participants viewed 4 sets of 2, 3, 4, or 5 stimuli.

The dependent variable for this task was participants’ score, computed as follows: Participants received 10 points if the first item is correctly recalled, 20 if both the first and second items are correctly recalled, 30 if the first three items are correctly recalled, and so on (each additional item in a series correctly recalled given that previous items were correctly recalled was worth 10 more points than the previous item). If an item in the series was forgotten, the scoring started over at 10 for the next correctly recalled item in the sequence.

![Figure 4. The Blockspan task.](image)

The second working memory task used in Study 2a was called Shapebuilder (Atkins et al., 2013). This task is similar to Blockspan but involved the tracking of additional information. Participants were asked to remember the order and spatial position in which a series of colored shapes were presented. They saw a 4 x 4 grid (see Figure 5) and a sequence of between two and four colored (red, blue, yellow, or green) shapes (circles, triangles, squares, or diamonds) appearing sequentially in one of the 16 possible grid locations. Participants were asked to remember the location, shape, and color of each item, and the order in which the items appeared. After the final item of a trial was presented, participants were asked to recreate the sequence by clicking on the correct colored shape and dragging it to the appropriate location. Participants completed 26 trials of length 2 to 4.
The Shapebuilder task increased in difficulty in two ways. First, trial length began at two and increased to three and then four. Second, within each set of trials of a given trial length, the trials became more difficult by including more diverse set of colors and shapes. At the easiest level, items were all the same shape or color, and at the most difficult level, items were all different colors and shapes. Participants saw the points awarded for each item immediately after releasing the mouse button.

Points were awarded as follows: participants received 15 points for the first item correctly recalled and additional 15 points for every consecutive item correctly recalled in the sequence. Shapebuilder also awarded points for partially recalled items; partial credit was only awarded when the correct location was guessed, such that the participant earned 5 points for the correctly recalled color but not shape, and 10 points for the correctly recalled shape but not color. Every time an item was missed, the scoring started over at 15 for the next correctly recalled item.

This task has shown good reliability in previous research: Scores on the even-numbered items correlated 0.63 with the odd numbered items, and both halves correlated with the total score at r>0.89 (Atkins et al., 2013). Additionally, Atkins et al. (2013) demonstrated that Shapebuilder scores correlated with two previously validated measures of working memory: Reading Span (Daneman & Carpenter, 1980) and Letter-Number Sequencing (Gold, Carpenter, Randolph, Goldberg, & Weinberger, 1997; Myerson, Emery, White, & Hale., 2003), as well as with a measure of visuo-spatial working memory called Blockspan (Atkins, 2011 as cited in Atkins et al., 2013).

**Procedure**

We followed the same general procedure as in Study 1. However, because the participants in Study 2a were native English speakers, they did not take the Versant™ Speaking Pro test or the LEAP-Q. In addition to the remaining tasks presented to participants in Study 1, participants in Study 2a completed the two working memory measures (Blockspan and Shapebuilder) after completing the listening comprehension portion of the study, and prior to being debriefed. Each working memory task required approximately ten minutes to complete.

**Data analysis**

The analyses treated length (short, medium, and long), and density (medium and high) as categorical variables, and working memory (WM) as a continuous variable. The dependent variables were multiple-choice comprehension score and recall score. We analyzed the data using mixed-effects (i.e., hierarchical) logistic regression, with length, density, and WM capacity as predictors of multiple-choice comprehension and recall accuracy (analyzed separately). The logistic regression model was similar to
those reported in the previous study, involving random participant and passage intercepts, and fixed effects for the factorial combination of the two experimentally manipulated factors of length (three levels: short, medium, long) and density (two levels: medium, high). In addition, WM was included as a covariate in a full factorial model. Details of the model can be found in Appendix H: Details of Bayesian statistical analysis.

Results

Demographic information

Participants’ ages ranged from 18 to 20, with an average age of 18.6 and a median age of 19. See Table I-3 in Appendix I for more detailed information on demographic descriptors.

Working memory

As the results on the comprehension questions and the recall items will be discussed with respect to the participants’ individual differences in working memory capacity, the working memory results are presented first. We discuss the multiple-choice response data next, and then turn to the free recall responses.

To ease interpretation of the model parameters, each participant’s working memory score was first converted to a z-score. This resulted in a relatively normal distribution of scores (see Figure 6).

![Histogram of working memory z-scores](image)

**Figure 6.** Distribution of working memory z-scores.

Multiple choice

To provide an overall view of performance on the multiple-choice questions, Table 5 shows the sample size and mean, SD, and minimum and maximum values of the percent correct for the multiple-choice data across participants. The mean number of items correct was quite high (~87%). Chance performance was at 25%. Even the participant with the lowest rate of accuracy performed well above chance.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>172</td>
<td>87.2%</td>
<td>9.4%</td>
<td>45.8%</td>
</tr>
</tbody>
</table>

**Table 5.** Descriptive statistics for the multiple-choice data included in the analysis for Study 2a.
Table 6 provides a summary of the mean percent correct for the multiple-choice comprehension questions across participants for each of the six conditions. As can be seen by comparing the values across the columns of Table 6, length did not have a consistent impact on accuracy: both long and short passages yielded superior performance compared to medium length passages. Comparing the values across the two rows also indicates that density did not consistently impact performance: accuracy was slightly higher for the medium density passages when these passages were of medium length, but accuracy was slightly higher for the high density passages when length was short or long. These results suggest that increasing density or length of the passages does not have a consistently negatively impact on listening comprehension as measured by the multiple-choice items.

Table 6. Mean accuracy (percent correct on the multiple-choice questions) and SD (in parentheses) as a function of passage length (short, medium, long).

<table>
<thead>
<tr>
<th>Density</th>
<th>Length</th>
<th>Short</th>
<th>Medium</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td></td>
<td>86.8%</td>
<td>(19.4%)</td>
<td>86.2%</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td>89.7%</td>
<td>(17.1%)</td>
<td>83.7%</td>
</tr>
</tbody>
</table>

Figure 7 shows the relationship between working memory capacity and the log-odds of a correct response to the multiple-choice items across the six conditions. As a reminder, because the dependent variable had a dichotomous outcome (i.e., 1 = correct, 0 = incorrect), our data analysis used a logistic regression in which performance was modeled in terms of the estimated log-odds of making a correct response—rather than the proportion correct—and the effect of a predictor indicated the corresponding change in log-odds (Gelman & Hill, 2007).

As can be seen, working memory seems to have a positive effect, such that individuals with greater working memory capacity have a higher probability of making a correct response. These effects appear similar across all six conditions, as indicated by the similar slopes of the lines. The impact of density on accuracy is, however, inconsistent across the three length conditions.

Figure 7. Accuracy (mean log-odds of making a correct response on the multiple-choice comprehension questions) as a function of passage length (short, medium, long), information density (medium, high), and working memory capacity.
These patterns illustrated in Table 6 and Figure 7 are supported by the logistic regression results; the model coefficients from the analysis are displayed in Figure 8.

![Multiple choice model results](image)

**Figure 8.** Model coefficients for the multiple-choice data.

Starting from the bottom row and moving up, the intercept at the bottom indicates that participants are performing quite well in the baseline reference condition, such that a participant with the sample mean working memory capacity would be expected to have a log-odds of success around 2.3, corresponding to around 90% accuracy.

Length did not have a significant effect on medium density passages, as indicated by the two simple effects of length, Length (med) (second from the bottom) and Length (long) (third label from the bottom), that have parameter values near zero and HDIs largely falling within the ROPE. Using the “divide-by-4” method of interpreting logistic regression coefficients (see the section on Model interpretation in Appendix H), a shift from short to medium or long passages would be expected to correspond to a maximum change in accuracy of 2.7% and 1.8%, respectively.

For short passages, higher information density is related to better performance, as indicated by the positive parameter of .274 that falls outside the ROPE and the HDI spanning [−0.01, 0.64], with 97% of the posterior distribution (i.e., credible values) being above zero. Finally, the significant negative parameter for the interaction between length (medium) and density ($\beta = -0.45$, HDI = [−1.03, −0.12]) indicates that the positive effect of density on short passages is negated and perhaps even reversed for medium passages. Indeed, we can compute the magnitude of the density effect for medium passages from the posterior distribution, and this indicates an effect of −0.20 with 89% of credible values falling below zero and an HDI spanning [−0.52, 0.08]. That is, density had a negative effect on medium passages,
although we are less certain about this effect than we are about the positive effect in short passages. In other words, the analysis supports the patterns noted above: length and density exact inconsistent effects on the accuracy of multiple-choice responses.

In summary, the varying combinations of length and density had an inconsistent impact on multiple-choice response performance. But it is not that the test did not reveal significant differences; the significant, simple effect of working memory on accuracy indicates that better working memory capacity is related to better performance for native speakers. For the five interaction terms involving working memory, all parameter estimates are near zero, and all HDIs contain zero, indicating that the effect of working memory does not differ across passage conditions. This was further supported by post-hoc pairwise comparisons of the slope estimates for each condition, which indicate that the HDIs for all comparisons included zero. Taken together, these findings suggest that for native speakers, working memory is a robust, positive predictor of performance for the multiple-choice response data, while passage length and density do not consistently predict performance on these items or interact with working memory.

Free recall

To provide an overall view of performance on the free recall questions, Table 7 shows the sample size and mean, SD, and minimum and maximum values of the percent correct for the free recall data across participants. Note that, as expected, and consistent with the results of Study 1, overall accuracy was much lower on the recall responses relative to the multiple-choice responses. Also, a greater range of performance was observed, with accuracy ranging from 20–90% correct.

Table 7. The number of participants included in the analysis of Study 2a, along with descriptive statistics for the proportion correct on recall questions, collapsed across all conditions.

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>172</td>
<td>63.1%</td>
<td>11.2%</td>
<td>20%</td>
<td>90%</td>
</tr>
</tbody>
</table>

Table 8 provides a summary of the mean percent correct across participants for each of the six conditions. The patterns in the table suggest that the effect of length on accuracy for the recall items appears to vary based on the density of the passage. For medium density passages, length appears to have a minimal effect on performance. In contrast, for high density passages, performance seems to drop steadily with increasing passage length. Note that this monotonic length effect with high density passages contrasts with the curvilinear relationship found with the multiple-choice data, where the worst performance was found in the medium length condition.

Table 8. Mean accuracy (percent correct on the free recall questions) and SD (in parentheses) as a function of passage length (short, medium, long) and information density (medium, high) as measured by type/token ratio.

<table>
<thead>
<tr>
<th>Density</th>
<th>Length</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short</td>
<td>Medium</td>
<td>Long</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>65.3%</td>
<td>(18.4%)</td>
<td>63.1%</td>
<td>(19.6%)</td>
<td>64.5%</td>
</tr>
<tr>
<td>High</td>
<td>66.4%</td>
<td>(18.5%)</td>
<td>61.6%</td>
<td>(16.7%)</td>
<td>58.2%</td>
</tr>
</tbody>
</table>

Figure 9 shows the relationship between working memory capacity and the log-odds of a correct response across the six conditions. Replicating the results from the multiple-choice data, working memory appears to have a positive relationship with performance on the free recall responses, such that individuals with greater working memory capacity have a higher probability of making a correct response. These
effects appear to be similar across all six conditions, as indicated by the similar slopes of the lines. In addition, the interaction between length and density on performance is very apparent in this figure. While the solid (medium density) and dotted (high density) lines for the short length condition lie nearly atop each other, the dotted line indicating performance on high density passages is a distance below the solid line in the medium length condition and is even further below in the long length condition. Moreover, the drop from short to medium length and medium to long length for the medium density passages is not as pronounced as that for the high density passages.

![Figure 9](image_url)

**Figure 9.** Accuracy (mean log-odds of making a correct response on the recall questions) as a function of passage length (short, medium, long), information density (medium, high), and working memory capacity.

These patterns are supported by the logistic regression results. The model coefficients from the analysis are displayed in Figure 10 below.
Starting from the bottom row and moving up, the intercept indicates that in the baseline reference condition a participant with average working memory capacity had a log-odds of making a correct response of .08, which corresponds to approximately 52% accuracy. No reliable change in performance was found for medium length passages, as indicated by the coefficient near zero ($\beta = -0.06$) and the HDI being contained entirely within the ROPE. Performance drops significantly for long length passages, however ($\beta = -0.16$, HDI $= [-0.34, -0.04]$). When switching to high density passages, there is no reliable change in performance on short passages ($\beta = 0.01$, HDI $= [-0.15, 0.14]$). The negative two-way interaction terms (Length(med) x Density(high) and Length(long) x Density(high)) indicate that performance may have dropped slightly going from short to medium length, high density passages ($\beta = -0.05$, HDI $= [-0.29, 0.14]$) but dropped significantly from short to long length, high density passages ($\beta = -0.27$, HDI $= [-0.46, -0.06]$). Note that for the interaction involving medium length, the parameter

---

[11] Note that this level of accuracy, even under the “lenient” scoring criteria applied, is quite impressive given that participants could have entered any English word (though the context provided by the recall item of course constrained the lexical class of the response).
estimate is negative but the HDI contains zero and only 69% of the most credible values fall below zero. This suggests that there is a fair amount of uncertainty regarding the difference between short and medium high density passages being a true negative effect. This stands in contrast to our certainty of the difference between short and long high density passages, where the HDI does not contain zero, so 100% of the most credible values fall below zero. Indeed, this was confirmed by post-hoc pairwise comparisons within the high density condition, which indicate marginal differences between short and medium length passages (Difference = 0.11, HDI = [−0.04, 0.26]) that were smaller in magnitude than the difference between short and long length passages (Difference = 0.43, HDI = [0.30, 0.57]) and between medium and long passages (Difference = .30, HDI = [.17, .47]). In other words, unlike with the multiple-choice data, the analysis of the recall questions indicates that the effect of density depends on the length of the passage; performance on medium density passages was similar to that for high passages when passages were short, was slightly worse for high passages when passages were of medium length, and was substantially (and significantly) worse when passages were long.

Looking at the parameter estimates involving working memory, we again saw that better working memory capacity is related to a higher probability of making a correct response in the baseline condition ($\beta = 0.10$, HDI = [0.02, 0.25]). The fact that 94% of the most credible values fall within the ROPE indicates that this effect is not large, despite being significantly different from zero. However, the effect is relatively stable across all six conditions, as indicated by the lack of any significant interaction terms involving working memory. Indeed, the HDIs for all working memory interactions contain zero and the HDIs of two of the interaction parameters are entirely contained within the ROPE, indicating that these interaction effects are practically indistinguishable from zero. This was further supported by post-hoc pairwise comparisons of the slope estimates for each condition, which indicate that the HDIs for all comparisons included zero. Thus, once again, working memory in and of itself was a robust, positive predictor of recall performance—replicating the results for multiple-choice items.
Discussion

The results of Study 2a are consistent with those for Study 1, finding no consistent impact of length on multiple-choice item performance. There was also not a consistent effect of information density on multiple-choice item performance. Short, high-density passages yielded superior performance to short, medium-density passages, but this pattern reversed for medium length passages, and there was no difference between performance for medium- and high-density long passages.

The results of Study 2a are also consistent with the Study 1 recall item results: recall item accuracy for long passages was significantly worse than for short passages. Further, the pattern of recall item responses indicates that passage length interacts with passage density. For short passages, medium- and high-density passages produced similar recall item performance, medium length passages of high-density produced somewhat worse recall performance than those of medium density, and for the long passages, accuracy on the recall items decreased substantially when density was high. While the impact of the density manipulation was not apparent for short passages, it was present for both medium and long passages, suggesting that the effect of increased density did not manifest until the passage was more difficult due to length.

Moreover, greater working memory capacity was related to more accurate comprehension on both multiple-choice and recall items, and the working memory effects were similar across all combinations of length and density. The fact that length and density effects were found even after controlling for differences in working memory suggests that the effects of length and density do not solely reflect increased working memory load, and that listening comprehension is a complex phenomenon that is impacted by all three factors.

STUDY 2B: EFFECTS OF PASSAGE LENGTH, INFORMATION DENSITY, AND WORKING MEMORY ON NON-NATIVE LISTENING COMPREHENSION

The results of Study 2a clearly indicate that even the high-density, long passages were not too difficult for native speakers of English. Although performance was not perfect on the multiple-choice and recall items across all levels of length and density, it was sufficiently high to suggest that variability in correct responses could be found with non-native listeners of English. Further, these results suggest that both length and density can influence listening comprehension for native listeners. It follows that the listening comprehension of non-native listeners should be even more impacted by increasing the level of these factors. We investigated this hypothesis in Study 2b.

Method

We designed Study 2b to determine how information density and passage length affect listening comprehension and recall accuracy for non-native speakers, and how these factors interact with each other. Like Study 2a, this study also explored the extent to which individual differences in working memory influence recall accuracy and listening comprehension under varying combinations of passage length and information density.

Participants

Study 2b included 188 native speakers of Spanish as participants. Of these, the data for 19 were excluded due to incomplete listening comprehension data (n = 1) or missing working memory data (n = 5). Data from an additional 13 participants were excluded due to performance at or below chance (i.e., 25% correct) on the multiple choice items, to prevent issues related to extremely low proficiency. The analysis was conducted on the remaining 169 participants (94 female, 72 male, 3 unknown). Of these, 88 participants were recruited in the Washington, D.C. metropolitan area through print ads in Spanish-language newspapers, fliers distributed or posted in public locations, and word of mouth via previous
study participants; the remaining 81 participants were recruited by a market research company in Austin, Texas.

Materials

This study was conducted using all of the same materials that were used in Study 2a (listening passages, comprehension questions, recall test items, and working memory tasks), but also including the Versant™ tests and LEAP-Q used in Study 1.

Listening passages

As in Study 2a, each participant heard a set of 12 passages comprised of six combinations of passage length (short, medium, and long) and density (medium and high). Two instances of each combination were presented, one at each ILR level (2 and 3). Every participant heard only one version of each passage, and the order in which the passages were presented to each participant was pseudo-randomized to avoid order effects.

Comprehension questions

The same comprehension items were used in this experiment as in Study 2a of the earlier investigation. However, in this case, Spanish translated versions were used instead of the original English versions.

Recall test

This experiment also used the exact same general recall items as in Study 2a. As in that study, the cued words differed for some density versions of a passage and the context provided for the cued word often differed between length versions of a passage.

Working memory capacity assessment

The same assessments of working memory capacity used in Study 2a were used in Study 2b.

Procedure

Participant screening

Native speakers of Spanish with an intermediate level of English listening proficiency were sought for participation in Study 2b. The screening procedure for identifying the targeted level of English listening proficiency was conducted over the phone and consisted of a brief conversation in English followed by a short self-assessment. The English conversation consisted of an overview of the research study and the collection of contact information. This exchange was used in conjunction with the results of the self-assessment to determine if the potential participant's English proficiency was at the targeted level. The researcher then administered a brief (5–10 minutes) self-assessment of English listening proficiency based on the ILR Self-Assessment of Foreign Language Listening Proficiency. The ILR self-assessment consists of Yes/No can-do statements for different language levels (from Level 1 to Level 5), does not include instructions, and presents all can-do statements as a single unit under each language level. The modified self-assessment used in the screening procedure included only Levels 1 to 3, contained instructions, presented each can-do statement separately irrespective of language level, and was translated into the language of the participants (Spanish).

Instructions for the CASL self-assessment were given in Spanish. Each statement was read to the potential participant and statements were repeated as necessary. The results were calculated by tallying all “Yes” responses. The initial eligible range for the targeted level (ILR Level 2) was 9–15 yes responses out of a total of 17 possible “Yes” responses. Later, potential participants were considered eligible even if they responded “Yes” to all 17 items if other aspects of their interaction with the researcher suggested they had Level 2 listening proficiency in English.
A market research subcontractor was also used to recruit participants. The screening procedure used by the market research firm was essentially the same as the CASL screening procedure, except for two modifications to the self-assessment instrument: the addition of a comprehension question and a scale for evaluating the comprehension of the question. The intent of the comprehension question was to reduce the subjectivity in the CASL screening procedure. An additional difference in the market research firm’s screening procedure was the inclusion of a question about computer literacy (ability to use a keyboard and a mouse).

**Experimental protocol**

For the main study, we followed the same general protocol as in Study 2a for native English-speaking participants, except the Spanish-speaking participants completed the Versant™ Speaking Pro test onsite at the time of the experimental session, which resulted in variations in the order in which the tasks were completed. Whenever possible, the Versant™ Speaking Pro test was completed at the beginning of the session. Due to the small number of available telephones relative to the number of participants to run within the experiment session, participants took the Versant™ Speaking Pro test at any interval between the tasks when a telephone was available. That is, any given participant could have taken the Versant™ Speaking Pro test at any interval before, between, or after the LEAPQ, the listening test, Blockspan, and Shapebuilder.

Upon arriving at the test site, participants provided informed consent in writing and then completed all experiment materials using an individual computer and headphones. A sample sound file was provided at the beginning of the study to allow participants to adjust the computer volume to a comfortable setting. All task instructions were presented in Spanish. Participants provided their unique identifier code which linked their Versant™ Speaking Pro test code, listening comprehension subscore, and working memory tasks to their responses. Participants completed the LEAP-Q and the portion of the experiment involving listening passages, comprehension questions, and recall items.

When administering the listening comprehension portion of the study, researchers adhered to the procedure outlined in the Defense Language Proficiency Testing System 5 Framework (2010) as closely as possible. The study included a short practice passage to familiarize participants with the procedure, followed by 12 experimental passages. For each passage, an orientation statement and the multiple-choice comprehension questions were presented on the screen while the passage played, and participants were allowed to answer the questions at any time. Each passage played twice without pausing and with a fixed interval between repetitions. Once participants answered the comprehension questions for that passage, they could press the “Next” button to proceed. On the following screen, the full set of recall items for that passage was presented. The passage audio was not available on the recall items screen. Participants typed their answers in English in the blanks and were allowed to answer in any order.

Participants had the option to take a roughly 10-minute break at the halfway mark. At the end of the study, participants were thanked, debriefed, and provided with instructions for accessing their Versant™ Speaking Pro test scores.

**Data analysis**

The analyses treated length (short, medium, and long), and density (medium and high) as categorical variables, and working memory capacity (WM) and Versant™ score as continuous variables. The dependent variables were multiple-choice comprehension score and recall score. We analyzed the data using mixed-effects (i.e., hierarchical) logistic regression, with length, density, working memory capacity, and L2 listening ability as predictors of multiple-choice comprehension and recall accuracy (analyzed separately). The logistic regression model was similar to those reported in the previous two studies, involving random participant and passage intercepts and fixed effects for the factorial combination of the two experimentally manipulated factors of length (three levels: short, medium, long) and density (two levels: medium, high). In addition, the model included WM and its interactions with the full factorial combination of the experimental factors. To control for differences in L2 listening ability, Versant™
scores were included as a simple covariate. Details of the model can be found in Appendix H: Details of Bayesian statistical analysis.

Results

Versant™ Pro Speaking test

The distribution of Versant™ listening comprehension subscores for participants in Study 2b is shown in Figure 11, below. As with participants in Study 1, some of the participants in Study 2b had scores outside the range corresponding to ILR 2/2+ (41–68), but the majority fell within this range. Versant™ scores were missing for 27 participants (roughly 16% of the sample) and were imputed within the Bayesian analysis. For a complete description of the imputation modeling procedure and an evaluation of its validity, see Appendix H: Details of Bayesian statistical analysis.

![Figure 11. Distribution of Versant™ listening comprehension subscores for participants in Study 2b.](image)

Demographic information

Eight participants did not respond to any of the LEAP-Q items. Of those who did respond, age ranged from 18 to 66 years, with an average age of 37.6 and a median age of 38. Roughly one-third of participants (40%) reported having a college degree, with a small proportion of those reporting post-baccalaureate education (12% of the entire group of participants). One percent of participants reported having less than high school. The median length of residence in a country where English is spoken was 11 years (range of 0–59 years of residence) and the median time spent in a work or school environment where English is spoken was 5 years (range of 0–38 years). See Tables I-4 and I-5 in Appendix I for more detailed information on demographic descriptors.

Working memory

As the results on the comprehension questions and the recall items will be discussed with respect to the participants’ individual differences in working memory capacity, the working memory results are presented first. We discuss the multiple-choice response data next and then turn to the free recall responses.
To ease interpretation of the model parameters, each participant’s working memory score was first converted to a z-score. This resulted in a relatively normal distribution of scores (see Figure 12).

![Figure 12. Distribution of working memory z-scores.](image)

As a means of comparing the participant groups from Study 2a and 2b, we conducted an independent t-test on the raw scores for working memory performance. The results demonstrate that native English speaker group had significantly higher working memory scores than the native Spanish speaker group both in the Blockspan task ($t[338.332] = 9.06, p < .001$) and the Shapebuilder task ($t[337.425] = 12.96, p < .001$). Because the working memory tasks were entirely visual and did not require the participants to use letters or words presented in their L2, they should be an accurate assessment of working memory. That said, it may be that the native Spanish speakers were less familiar with computers, and that their lack of facility with the mouse had an impact on their scores.

**Multiple choice**

To provide an overall view of performance on the multiple-choice questions, Table 9 shows the sample size and mean, $SD$, and minimum and maximum values of the percentage correct for the multiple-choice data across participants. Chance performance was at .250. All participants responded above chance levels.

**Table 9.** Descriptive statistics for the multiple-choice data included in the analysis for Study 2b.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>169</td>
<td>59.6%</td>
<td>15.5%</td>
<td>91.7%</td>
</tr>
</tbody>
</table>

Table 10 provides a summary of the mean percent correct across participants for each of the six conditions. The patterns in the table suggest that information density had minimal impact on performance across all three passage lengths. With respect to passage length, medium length passages were the most difficult passages. It is unclear why these passages were harder than their short and long counterparts. It

---

12 Because we cannot assume the two populations have equal variances, the degrees of freedom for the t-tests were calculated using the Welch approximation.
may be that there are two factors at play: there was less distracting information in the short passages and the long passages gave the participants more time to process what they were hearing so they could answer the questions more accurately.

Table 10. Mean accuracy (percent correct on the multiple-choice questions) and SD (in parentheses) as a function of passage length (short, medium, long) and density (medium and high).

<table>
<thead>
<tr>
<th>Density</th>
<th>Length</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Short</td>
<td>Medium</td>
<td>Long</td>
</tr>
<tr>
<td>Medium</td>
<td>61.4%</td>
<td>(25.1%)</td>
<td>56.5%</td>
</tr>
<tr>
<td>High</td>
<td>61.4%</td>
<td>(24.2%)</td>
<td>57.7%</td>
</tr>
</tbody>
</table>

Figure 13 shows the relationship between working memory capacity and the log-odds of a correct response across the six conditions, as estimated from the logistic regression analysis in which individual differences in L2 listening ability (i.e., Versant™ scores) were also taken into account. Working memory has a small but positive impact on accuracy, such that individuals with greater working memory capacity are slightly more likely to make a correct response. This relationship appears to be attenuated somewhat in the medium length, medium density condition, as indicated in the middle panel of Figure 13 by the solid line that is less steep than the dotted line. The two density conditions yield nearly identical log-odds for each level of length, and although there is a decrease in accuracy between the short and medium length conditions, the long length condition log-odds are higher than those in the medium length condition.

![Figure 13](image)

**Figure 13.** Accuracy (mean log-odds of making a correct response on the multiple-choice comprehension questions) as a function of passage length (short, medium, long), information density (medium, high), and working memory capacity, while controlling for differences in L2 listening ability.

These patterns are supported by the logistic regression results. The model coefficients from the analysis are displayed in Figure 14.
Figure 14. Model coefficients for the multiple-choice data, while controlling for differences in L2 listening ability.

Starting from the bottom row and moving up, the intercept at the bottom indicates that in the baseline reference condition, a non-native speaker with ILR Level 2 listening ability and with the sample mean working memory capacity would be expected to have a log-odds of success around 0.71, corresponding to around 67% accuracy. For medium density passages, medium length is related to lower accuracy, as indicated by the negative Length (med) parameter value of $-0.21$ and an HDI spanning $[-0.43, -0.01]$, with 97% of the posterior distribution (i.e., credible values) falling below zero. In contrast, long length had no effect on accuracy when compared to short length passages, as indicated by the Length (long) coefficient near zero ($\beta = -0.02$) and an HDI that was centered near zero and had nearly 89% of its values contained within the ROPE. Indeed, using the “divide-by-4” method of interpreting logistic regression coefficients (see the Model interpretation section in Appendix H), a shift from short to long passages would be expected to correspond to a maximum change in accuracy of merely 0.6%. For short passages, information density had no effect on performance ($\beta < 0.01$, HDI = $[-0.22, 0.21]$); no other significant interactions were found between length and information density. In other words, the analysis indicates that there was no consistent effect of either length or density on multiple-choice item performance; accuracy was lowest in the medium length conditions and did not differ with density level.

The simple effect of working memory indicates that better working memory capacity is related to better performance on short, medium density passages ($\beta = .19$); although the HDI $[-0.03, 0.38]$ contains zero, the fact that 95.5% of the most credible values fall above zero suggests a reasonable amount of confidence that this positive effect is stable. The negative interaction between working
memory in medium length passages appears to be stable, as indicated by the parameter estimate ($\beta = -0.17, \text{HDI} = [-0.43, 0.09]$) and the fact that over 92% of the credible values fall below zero. In other words, the working memory slope was attenuated for medium length, medium density passages. For the remaining working memory interaction parameters, all parameter estimates are near zero, and all HDIs contained (and were centered near) zero, indicating that the effect of working memory does not differ across these other passage conditions (i.e., the slopes were similar to those observed in the short, medium density passages). This was further supported by post-hoc pairwise comparisons of the slope estimates for each condition, which indicated a smaller working memory slope in the medium length, medium density passages relative to short length passages (regardless of information density). Taken together, these findings suggest that working memory is a robust, positive predictor of performance for the multiple-choice response data, although the working memory effect may be somewhat attenuated in the most difficult condition where passages were medium length and had medium information density.

It is noteworthy that this effect of working memory was found even when controlling for individual differences in L2 listening ability. In a separate analysis in which Versant™ scores were not included in the model, the working memory effect in the baseline condition was much stronger ($\beta = 0.39, \text{HDI} = [0.18, 0.58]$), as was the WM x Length(medium) interaction ($\beta = -0.19, \text{HDI} = [-0.52, 0.05], 91\%$ of the posterior below zero). Recall that we purposefully selected our two measures of working memory to be relatively language independent (i.e., they did not involve the processing or storage of L1 or L2 words, sentences, etc.), in order to reduce any potential confound between the working memory measures and L2 proficiency. The fact that the working memory effect is attenuated but still present after controlling for differences in L2 listening ability suggests that working memory is related both to global L2 listening proficiency (as measured by the Versant™ listening test) and to the L2 listening abilities measured by our listening task (for similar claims of the importance of working memory to both global and specific L2 abilities, see Linck, Osthus, Koeth, & Bunting, 2012; 2013).

**Free recall**

To provide an overall view of performance on the free recall questions, Table 11 shows the sample size and mean, SD, and minimum and maximum values of the proportion correct for the free recall data across participants. Note that, as for the native English speakers in Study 2a, overall accuracy was much lower on the recall responses relative to the multiple-choice responses. Also, a greater range of performance was observed, with accuracy ranging from 20–90% correct.

**Table 11.** The number of participants included in the analysis of Study 2b, along with descriptive statistics for the proportion correct on recall questions, collapsed across all conditions.

<table>
<thead>
<tr>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>169</td>
<td>32.7%</td>
<td>20.6%</td>
<td>0%</td>
<td>90%</td>
</tr>
</tbody>
</table>

Table 12 provides a summary of the mean percent correct across participants for each of the six conditions. Unlike the multiple-choice data, there appears to be an overall effect of length, such that long passages are more difficult than short passages. There also appears to be an overall effect of the information density of the passage, with high density passages being more difficult than medium density passages. Note that this monotonic length effect contrasts with the curvilinear relationship found with the multiple-choice data, where the worst performance was found in the medium length condition.
Table 12. Mean (and SD) accuracy (percent correct on the free recall questions) as a function of passage length (short, medium, long).

<table>
<thead>
<tr>
<th>Density</th>
<th>Short</th>
<th>Medium</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>35.1% (25.8%)</td>
<td>35.2% (27.2%)</td>
<td>32.7% (23.4%)</td>
</tr>
<tr>
<td>High</td>
<td>32.7% (24.2%)</td>
<td>31.3% (24.4%)</td>
<td>29.9% (23.7%)</td>
</tr>
</tbody>
</table>

Figure 15 shows the relationship between working memory capacity and the log-odds of a correct response across the six conditions, as estimated from the logistic regression analysis in which we controlled for individual differences in L2 listening comprehension. As was true for the multiple-choice items, working memory has a positive, but smaller, impact on performance on the free recall responses, such that individuals with greater working memory capacity have a higher probability of making a correct response. These effects appear to be similar across all six conditions, as indicated by the similar slopes of the lines.

The impact of density on accuracy for the recall items appears to be consistent across the three levels of length (roughly the same distance between the solid and dotted lines at each level of length), and the effect of length holds for both levels of density (the same size decrease in accuracy at each increasing level of length for both the solid and the dotted lines).

Figure 15. Accuracy (mean log-odds of making a correct response on the recall questions) as a function of passage length (short, medium, long), information density (medium, high), and working memory capacity, while controlling for differences in L2 listening ability.

These patterns are supported by the logistic regression results. The model coefficients from the analysis are displayed in Figure 16.
Starting from the bottom row and moving up, the intercept indicates that, in the baseline reference condition, a participant with ILR Level 2 listening ability and with average working memory capacity had a log-odds of making a correct response of −0.34, which corresponds to approximately 42% accuracy. For medium density passages, no detectable change was found for medium passages, as indicated by the coefficient near zero (β = −0.01) and the HDI being contained entirely within the ROPE. However, performance drops for long length passages (β = −0.17, HDI = [−0.36, −0.01]).

The negative effect of density (β = −0.15, HDI = [−0.33, 0.02], over 95% of credible values below zero) indicates that performance was worse on high density passages. The near-zero parameters for the two interactions between length and density indicate negligible effects (βs = −0.07 and less than −0.01, respectively). In other words, the analysis of the recall questions indicates that length and density independently affected performance (i.e., did not interact).

Looking at the parameter estimates involving working memory, we see that working memory capacity is not significantly related to performance in the baseline condition (β = 0.08, HDI = [−0.12, 0.28]), and 86% of the most credible values fall inside the ROPE. Moreover, there are no significant interactions, indicating that no real effect of working memory was detected across all six conditions. However, it is important to note that we controlled for individual differences in L2 listening ability in this
analysis. In a separate analysis in which Versant™ scores were not included in the model, working memory was indeed a strong, positive predictor of recall accuracy ($\beta = 0.57$, HDI = [0.35, 0.80]). Here, unlike the multiple-choice results, the lack of a working memory effect on recall accuracy when L2 listening ability is accounted for in the analysis suggests that the more open-ended free recall outcome measure was perhaps sufficiently difficult for these L2 learners to eliminate the benefits of having greater working memory capacity.

Discussion

The results of Study 2b indicate that passage length and information density do not consistently impact listening comprehension performance for non-native listeners when comprehension is measured with multiple-choice items. Multiple-choice item accuracy was lower for medium length passages but was similar for short and long passages and did not differ with density level. Length and density both had significant effects on recall task performance, however. Recall performance was lower when passages were dense, and this effect was consistent across all three passage lengths. Length also impacted recall accuracy such that performance for long passages was worse than that for short passages for both medium and high density passages (though the effect was significant only for medium density passages). It is important to consider that the higher average age of Study 2b participants compared to Study 2a participants may suggest lower overall computer literacy skills, which could be a contributing factor to the results observed. Nevertheless, English listening proficiency strongly predicted performance on both tasks, consistent with the results of Study 1. In contrast to the results of Study 2a, greater working memory capacity was related to more accurate responses for only multiple-choice items. A separate analysis indicated that working memory’s effect on performance was attenuated (to the extent of being non-significant for recall items) when English listening proficiency was included in the model. This suggests that working memory is related both to global L2 listening proficiency and to the L2 listening abilities measured by our listening task (for similar claims of the importance of working memory to both global and specific L2 abilities, see Linck, Osthus, Koeth, & Bunting, 2012; 2013).

CONCLUSIONS

The results of Studies 1 and 2b reveal that even dramatic manipulations of the length of listening passages (from ~1 minute to ~3 minutes) fail to impact the performance of non-native listeners on multiple-choice comprehension items in a systematic way: although participants were less accurate in answering items for medium length passages than for short length passages, they were more accurate in answering items for long length passages than short length passages. It is important to note that the lack of a consistent effect of length for non-native listeners was not the result of a floor effect: overall accuracy for the multiple-choice items was nearly 60%, well above chance (25%). The results of Study 2a also fail to show a consistent impact of length on multiple-choice item accuracy for native listeners, revealing the same pattern of results: items for medium length passages were harder to answer than those for both short and long passages.

The lack of an interpretable and consistent effect of length for multiple-choice items is perhaps not surprising in light of the fact that the multiple choice items were available before, during, and after the corresponding passage played. The availability of the items throughout passage play meant that the participants did not need to retain the information necessary for answering the questions until the completion of the passage and could in fact answer items during the first play of the passage and check their answers when the passage replayed. Under these circumstances, increasing the length of the passage did not increase the amount of information the listener had to retain until answering the multiple-choice questions. The non-linear impact of passage length on performance for the multiple-choice items is somewhat difficult to interpret, but one potential explanation is that the medium length passages contained enough additional material to confuse or overwhelm listeners but not enough to provide the benefit of additional context or enough time to allow listeners to process the information necessary for the
multiple-choice items. The results for multiple-choice items overall suggest that the impact of length on listening comprehension is not straight-forward for either native or non-native listeners.

In contrast to the multiple-choice items, there was a small but consistent negative effect of length on recall item performance in Study 1, and a more sizable negative effect in Study 2b. Even for native listeners, longer passages yielded worse performance on the recall items (in Study 2a), though this effect was most pronounced when passages were especially difficult due to having high density. This pattern of results suggests that the need to process and retain a greater amount of information until moving on to the recall items negatively affected recall of the specific words heard. Unlike the multiple-choice items, the recall items were not available to listeners until they had finished hearing all listening material and had answered the multiple-choice items. For this reason, these items were likely to make greater demands for listeners to retain the entirety of the passage information for at least a short period of time. The differential results of the length manipulations on recall and multiple-choice items are also consistent with previous findings for native speakers in which passage length had an impact on recall but not on recognition (Schultz, Jr. & Johnson, 1982). Since the multiple-choice items presented potential answers to participants, recognition of previously heard information was a component of this task. In contrast, recall items were open-ended, so this task did not involve recognition.

In addition to examining how length affected performance on multiple-choice and recall comprehension items, Studies 2a and 2b investigated the impact of information density (type/token ratio) and its potential interaction with length; these studies also explored how participants’ working memory capacity predicted performance. As for length, neither study found a significant impact of information density on multiple-choice item performance. However, the results for the recall items suggest that increased information density does create difficulty for listening comprehension. The native listeners in Study 2a showed an effect of density on recall item performance that was dependent on length: only for the long passages did performance suffer on the high compared to the medium density passages. This result parallels that for length on the recall items, in that the impact of a difficulty manipulation is most apparent for native listeners when a passage is difficult due to other factors (long length, in this case). Similarly, performance on the recall items suffered for non-native listener participants in Study 2b when density of the passage was high. For these participants, however, high density passages were more difficult than medium density passages even when passage length was short. This pattern of results indicates that for non-native listeners even the short length passages are sufficiently difficult to see an impact of increasing density. Combined, the results for the native and non-native listeners support a role for density and passage length in listening comprehension as measured through recall.

Across the three studies, the most consistent factors predicting performance were English listening proficiency (for non-native listeners in Studies 1 and 2b) and working memory capacity (in Studies 2a and 2b). Working memory capacity had a consistent positive impact on recall and multiple-choice item performance for native speakers; this effect did not differ depending on the length or density of the passage. Working memory also predicted performance on multiple-choice and recall items for non-native listeners, though this effect was attenuated by including participants’ English listening proficiency in the model, a factor which also positively predicted performance for these participants. Working memory may have played a role both in the general skills tapped by the English proficiency test and in the specific skills required to successfully complete the multiple-choice and recall tasks, resulting in a situation where the variance accounted for by differences in working memory capacity is partially overlapping with that accounted for by differences in English listening proficiency.

In sum, the results of the three studies reported here suggest that length and density do not consistently impact performance on multiple-choice comprehension items of the type included on the DLPT5, but do influence listening comprehension difficulty for other tasks, such as recall of words used in the passage. In addition, the results of Studies 1, 2a, and 2b indicate that increased information density does not interfere with comprehension as measured by multiple-choice items presented co-temporally with the listening passage.
REFERENCES


APPENDIX A: SAMPLE PASSAGE (MEDIUM LENGTH, LOW TYPE/TOKEN RATIO)

JB: This is Earth and Sky with a survivor's story.

DB: In the past fifty years, about four million coyotes have been killed in the United States. And these coyotes have been killed not just by ranchers and farmers alone. They’ve been killed by government agencies. In the United States, a government agency called Wildlife Services killed eighty-six thousand coyotes in just the year nineteen ninety-nine alone. But the story of the coyote is a survivor’s story.

JB: Many years ago, coyotes were known to live just in the western United States, but now they are known to live in almost every state in the United States. They’re known to be in the Bronx! Dr. Bekoff from the University of Colorado has been studying coyotes for more than twenty-five years. He told us how the coyote can live through the kind of persecution that pushed other animals to extinction.

Dr. Bekoff: Coyotes are animals with a lot of adaptability. They can live in deserts or they can live in mountains. They can live in cold or they can live in warmth, and they can live on mice, uh they can live on lizards. Coyotes have been known to eat rubber, coyotes have been known to eat clothing.

DB: Coyotes are learning to live in a human-dominated world. And Dr. Bekoff says we can learn a lot about adaptability from coyotes by letting them be.
### APPENDIX B: LENGTH OF PASSAGES (NUMBER OF SYLLABLES)

<table>
<thead>
<tr>
<th>Base Level</th>
<th>Short Title</th>
<th>Short Length (213–243 syllables)</th>
<th>Medium Length (328–406 syllables)</th>
<th>Long Length (699–773 syllables)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Low Density</td>
<td>Medium Density</td>
<td>High Density</td>
</tr>
<tr>
<td>2</td>
<td>Refugees</td>
<td>224</td>
<td>224</td>
<td>231</td>
</tr>
<tr>
<td></td>
<td>Vitamin D</td>
<td>213</td>
<td>214</td>
<td>221</td>
</tr>
<tr>
<td></td>
<td>***Coyote</td>
<td>219</td>
<td>223</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>AM/PM</td>
<td>243</td>
<td>243</td>
<td>236</td>
</tr>
<tr>
<td></td>
<td>Taser</td>
<td>213</td>
<td>218</td>
<td>218</td>
</tr>
<tr>
<td></td>
<td>Road and Rail</td>
<td>230</td>
<td>217</td>
<td>226</td>
</tr>
<tr>
<td>3</td>
<td>***Cosby</td>
<td>222</td>
<td>223</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>*Charter Schools</td>
<td>214</td>
<td>220</td>
<td>229</td>
</tr>
<tr>
<td></td>
<td>***Secretary Rice</td>
<td>225</td>
<td>218</td>
<td>222</td>
</tr>
<tr>
<td></td>
<td>Kurds</td>
<td>226</td>
<td>233</td>
<td>224</td>
</tr>
<tr>
<td></td>
<td>*Terrorists</td>
<td>227</td>
<td>224</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>*Artful Brain</td>
<td>224</td>
<td>225</td>
<td>225</td>
</tr>
</tbody>
</table>

* Single speaker, *** three speakers. All remaining passages have two speakers.
## APPENDIX C: DENSITY OF PASSAGES (TYPE/TOKEN RATIO CALCULATED USING MATTR)

<table>
<thead>
<tr>
<th>Base Level</th>
<th>Short Title</th>
<th>Short Length</th>
<th>Medium Length</th>
<th>Long Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Density (0.49–0.54)</td>
<td>Medium Density (0.59–0.63)</td>
<td>High Density (0.70–0.74)</td>
<td>Low Density (0.49–0.55)</td>
<td>Medium Density (0.59–0.65)</td>
</tr>
<tr>
<td>2</td>
<td>Refugees</td>
<td>0.53</td>
<td>0.59</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Vitamin D</td>
<td>0.53</td>
<td>0.61</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>***Coyote</td>
<td>0.54</td>
<td>0.63</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>AM/PM</td>
<td>0.50</td>
<td>0.63</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>Taser</td>
<td>0.55</td>
<td>0.62</td>
<td>0.71</td>
</tr>
<tr>
<td></td>
<td>Road and Rail</td>
<td>0.50</td>
<td>0.60</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>***Cosby</td>
<td>0.54</td>
<td>0.63</td>
<td>0.72</td>
</tr>
<tr>
<td></td>
<td>*Charter Schools</td>
<td>0.49</td>
<td>0.62</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>***Secretary Rice</td>
<td>0.53</td>
<td>0.62</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>Kurds</td>
<td>0.49</td>
<td>0.59</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>*Terrorists</td>
<td>0.54</td>
<td>0.61</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>*Artful Brain</td>
<td>0.52</td>
<td>0.63</td>
<td>0.74</td>
</tr>
</tbody>
</table>

* Single speaker, *** three speakers. All remaining passages have two speakers.
APPENDIX D: SAMPLE MULTIPLE-CHOICE COMPREHENSION QUESTIONS

1. According to the report, what has occurred with coyotes during the last several decades?

They have expanded the area where they live.*
They have attacked more ranch animals.
They have moved away from farm areas.
They have been removed from western states.

2. According to the researcher, Dr. Bekoff, how have coyotes been able to thrive?

Coyotes are able to adapt.*
Coyotes have few enemies.
Coyotes are very intelligent.
Coyotes live far from people.
## APPENDIX E: SAMPLE RECALL ITEMS

### Medium type/token ratio

<table>
<thead>
<tr>
<th>Keyword: <em>pushed</em></th>
<th>Short: He told us how the coyote is able to live through the kind of persecution that other animals toward extinction.</th>
<th>Medium: He told us how the coyote can live through the kind of persecution that the bison, wolf, and passenger pigeon to extinction.</th>
<th>Long: Dr. Bekoff told us how the coyote can survive the kind of persecution that the bison, wolf, and passenger pigeon to extinction.</th>
</tr>
</thead>
</table>

### High type/token ratio

<table>
<thead>
<tr>
<th>Keyword: <em>pushed</em></th>
<th>Short: He told us how this animal is able to survive the kind of persecution that other species toward extinction.</th>
<th>Medium: He told us how the coyote is able to survive the kind of persecution that the bison, wolf, and passenger pigeon toward extinction.</th>
<th>Long: He told us how the coyote is able to survive the kind of persecution that the bison, wolf, and passenger pigeon toward extinction.</th>
</tr>
</thead>
</table>
APPENDIX F: MODIFIED VERSION OF THE LEAP-Q

Based on LEAP-Q from Marian, Blumenfeld, & Kaushanskaya (2007).

LANGUAGE EXPERIENCE AND PROFICIENCY QUESTIONNAIRE (LEAP-Q)

[Page 1]

Enter your 8-digit participant ID in the square below: ______________________

Today’s Date: ______________________
Age: ______________________
Date of Birth: ______________________

Please list all the languages you know in order of dominance:
1: ______________________
2: ______________________
3: ______________________
4: ______________________
5: ______________________

Please list all the languages you know in order of acquisition (your native language first):
1: ______________________
2: ______________________
3: ______________________
4: ______________________
5: ______________________

How many years of formal education do you have? ______

Please check your highest education level (or the approximate US equivalent to a degree obtained in another country):

__ Less than High School
__ Some College
__ Masters
__ High School
__ College
__ Ph.D./M.D./J.D.
__ Professional Training
__ Some Graduate School
__ Other

(9) Have you ever had: a vision problem , hearing impairment , language disability , or learning disability (Check all applicable). If yes, please explain (including any corrections):
__ a vision problem: ______________________
__ hearing impairment: ______________________
__ language disability: ______________________
This is my [native/second/third/fourth/fifth] language. All questions below refer to your knowledge of X language.

(1) Age when you:
   ...began acquiring X: _______
   ...became fluent in X: _______
   ...began reading in X: _______
   ...became fluent reading in X: _______

(2) Please list the number of years and months you spent in each language environment:
   A country where X is spoken: ______ years ______ months
   A family where X is spoken: ______ years ______ months
   A school and/or working environment where X is spoken: ______ years ______ months

(3) On a scale from zero to ten, please select your level of proficiency in speaking, understanding, and reading X from the scroll-down menus:
   Scale: 0 = none, 1 = very low, 2 = low, 3 = passable, 4 = below average, 5 = average, 6 = above average, 7 = good, 8 = very good, 9 = excellent, 10 = perfect

<table>
<thead>
<tr>
<th>Speaking</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Understanding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>spoken language</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(4) On a scale from zero to ten, please select how much the following factors contributed to you learning X:
   Scale: 0 = not at all, 5 = moderate contribution, 10 = most important contribution

<table>
<thead>
<tr>
<th>Interacting with friends</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language tapes/self</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instruction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interacting with family</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watching TV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening to the radio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(5) Please rate to what extent you are currently exposed to X in the following contexts:
   Scale: 0 = not at all, 5 = half the time, 10 = all the time

<table>
<thead>
<tr>
<th>Interacting with friends</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language tapes/self</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>instruction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interacting with family</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Watching TV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening to the radio</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
APPENDIX G: ILR SELF-ASSESSMENT OF FOREIGN LANGUAGE LISTENING PROFICIENCY

This self-assessment was developed based on the language proficiency levels set forth by the Interagency Language Roundtable.

SELF-ASSESSMENT OF FOREIGN LANGUAGE LISTENING PROFICIENCY

To estimate your level of proficiency, start at the lowest level and respond to each statement. For each statement, respond “yes” or “no.” If a statement describes your ability only some of the time or only in some contexts, you should answer “no.” If you answer “yes” to every statement in the level, your ability is probably at least at that level. Move on to the descriptions at the next level. If you answer “no” to one or more statements, then you are likely not at that level.

<table>
<thead>
<tr>
<th>Level 1</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>In everyday conversation with people speaking the standard dialect, I can understand speech that is slow and clear.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can understand basic directions and instructions, such as how to get to a local store.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can understand questions and answers about basic survival needs, such as meals, lodging, transportation and time.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can understand routine questions about my job, my immediate family and myself.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can understand simple statements about a person’s background and occupation.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>If I cannot understand what a speaker tells me, I can understand the statement after it has been repeated or rephrased slowly and clearly.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>When people are speaking the standard dialect at a normal rate, I can understand their speech when it is spoken with some repetition and rephrasing, can understand speech about everyday topics, for example common personal and family news, well-known current events, and routine situations at work.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can understand spoken descriptions of different places, for instance the geography of a country or location that is familiar.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can understand uncomplicated stories about current, past and future events.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can understand at least some details from announcements made over a loudspeaker.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can usually understand the main idea and basic facts from a short news report on the radio or television.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>I can accurately follow all conversations among native speakers who are speaking at a normal rate of speech.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I rarely, if ever, have to ask speakers to paraphrase or explain what they have said.</td>
<td>☐ Yes ☐ No</td>
<td></td>
</tr>
<tr>
<td>I can correctly infer meanings that are not directly stated.</td>
<td>□</td>
<td>Yes</td>
</tr>
<tr>
<td>-------------------------------------------------------------</td>
<td>----</td>
<td>-----</td>
</tr>
<tr>
<td>I can understand discussions of ideas and concepts, including proposals and speculation.</td>
<td>□</td>
<td>Yes</td>
</tr>
<tr>
<td>I can understand someone’s opinion and the points used to support the opinion.</td>
<td>□</td>
<td>Yes</td>
</tr>
<tr>
<td>I can often, if not always, detect the attitudes and feelings of a speaker.</td>
<td>□</td>
<td>Yes</td>
</tr>
<tr>
<td>I can understand speech in a professional setting concerning my field of expertise or some technical subjects, such as a lecture or a panel discussion.</td>
<td>□</td>
<td>Yes</td>
</tr>
</tbody>
</table>
APPENDIX H: DETAILS OF BAYESIAN STATISTICAL ANALYSIS

The full factorial models used for analyzing the data for each study are shown below. Note the following differences among the models: Studies 1 and 2b, but not Study 2a, included Versant™ scores as a simple covariate; and for Studies 2a and 2b, but not Study 1, the models included the factorial combination of length (medium, long), density (medium, high), and Working Memory scores. See below for the equations denoting the regression models for each study. Following the equations, we describe the models and their respective interpretations.

Study 1 (Passage Length, controlling for L2 Listening Ability)

\[ \ln \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \beta_0 + u_i + u_j + \beta_1 \text{length}(\text{medium})_{ij} + \beta_2 \text{length}(\text{long})_{ij} + \beta_3 \text{Versant}_i \]

Study 2a (Passage Length, Information Density, and Working Memory Capacity)

\[ \ln \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \]
\[ \beta_0 + u_i + u_j + \beta_1 \text{length}(\text{medium})_{ij} + \beta_2 \text{length}(\text{long})_{ij} + \beta_3 \text{density}(\text{high})_{ij} + \beta_4 \text{length}(\text{medium})_{ij} \text{density}(\text{high})_{ij} + \beta_5 \text{length}(\text{high})_{ij} \text{density}(\text{high})_{ij} + \beta_6 \text{WM}_i + \beta_7 \text{length}(\text{medium})_{ij} \text{WM}_i + \beta_8 \text{length}(\text{long})_{ij} \text{WM}_i + \beta_9 \text{density}(\text{high})_{ij} \text{WM}_i + \beta_{10} \text{length}(\text{medium})_{ij} \text{density}(\text{high})_{ij} \text{WM}_i + \beta_{11} \text{length}(\text{high})_{ij} \text{density}(\text{high})_{ij} \text{WM}_i + \beta_{12} \text{Versant}_i \]

Study 2b (Passage Length, Information Density, and Working Memory Capacity, controlling for L2 Listening Ability)

\[ \ln \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \]
\[ \beta_0 + u_i + u_j + \beta_1 \text{length}(\text{medium})_{ij} + \beta_2 \text{length}(\text{long})_{ij} + \beta_3 \text{density}(\text{high})_{ij} + \beta_4 \text{length}(\text{medium})_{ij} \text{density}(\text{high})_{ij} + \beta_5 \text{length}(\text{high})_{ij} \text{density}(\text{high})_{ij} + \beta_6 \text{WM}_i + \beta_7 \text{length}(\text{medium})_{ij} \text{WM}_i + \beta_8 \text{length}(\text{long})_{ij} \text{WM}_i + \beta_9 \text{density}(\text{high})_{ij} \text{WM}_i + \beta_{10} \text{length}(\text{medium})_{ij} \text{density}(\text{high})_{ij} \text{WM}_i + \beta_{11} \text{length}(\text{high})_{ij} \text{density}(\text{high})_{ij} \text{WM}_i + \beta_{12} \text{Versant}_i \]

In all of the models, \( p_{ij} \) is the modeled probability of a correct response for participant \( i \) and passage \( j \); \( \beta_0 \) is the intercept; \( u_i \) is the subject-specific adjustment to the intercept for participant \( i \); \( u_j \) is the passage-specific adjustment to the intercept for passage \( j \); and \( \text{length}(\text{medium})_{ij} \) and \( \text{length}(\text{long})_{ij} \) are binary indicator variables (1 = yes, 0 = no) for a medium and long length, respectively, for passage \( j \) heard by participant \( i \). In the models for Studies 1 and 2b, Versant\textsuperscript{™} is the Versant\textsuperscript{™} score for participant \( i \). In the models for Studies 2a and 2b, \( \text{density}(\text{high})_{ij} \) is a binary indicator variable for high density for passage \( j \) heard by participant \( i \), and \( \text{WM}_i \) is the working memory z-score for participant \( i \).

Due to the dummy coding for length and density, and the use of standardized working memory scores (in Studies 2a and 2b) and standardized Versant\textsuperscript{™} scores (in Studies 1 and 2b), the interpretation of the intercept parameter varies slightly across the three studies. For Study 1, the intercept corresponds to the overall propensity of the average participant with the sample average L2 listening comprehension ability to give correct responses to questions from the average short passage. For Study 2a, the intercept corresponds to the overall propensity of the average participant with the sample average working memory capacity to give correct responses to questions.
from the average short, medium density passage. For Study 2b, the intercept corresponds to the overall propensity of the average participant with the sample average working memory capacity and L2 listening comprehension ability to give correct responses to questions from the average short, medium density passage. Due to the logistic function, a significantly positive intercept would indicate that the probability of making a correct response is higher than 50%, a significantly negative intercept would indicate that the probability of making a correct response is lower than 50%, and a non-significant intercept at or near 0 would indicate that the probability of making a correct response is at or near 50%.

In all of the models, the parameters for length and density indicate how the probability of correctly responding increases or decreases with changes in length and density, respectively. When included, the parameter for working memory indicates how the probability of correctly responding changes with a one standard deviation change in working memory; similarly, the parameter for L2 listening ability (when included) indicates how the probability of correctly responding changes with a one standard deviation change in Versant™ score. In the models for Studies 2a and 2b, the parameters for the interaction terms involving two categorical factors (e.g., $\beta_4$ in Study 2a) indicate how the probability of a correct response changes with particular combinations of factors. In this way, these parameters indicate contrasts (or deviations) from the baseline levels for length and density. However, the parameters for the interaction terms involving working memory indicate how the slope for working memory (i.e., $\beta_6$ in Studies 2a and 2b) changes across particular combinations of factors. That is, these parameters indicate adjustments to the baseline condition working memory slope.

**Specification of prior distributions**

As discussed above in the body of the report (see *Prior Distributions*), we set weakly informative priors for all model parameters to constrain the model to make unreasonably extreme parameter values less likely *a priori*, while still letting the data be the primary source of information driving the values observed in the posterior distribution (Kruschke, 2011). In WinBUGS, for normal distributions, prior distributions are parameterized by mean and precision, rather than SD, with precision defined as the inverse of the variance (Kruschke, 2011). For the intercept and all beta parameters, the priors were set as a normal distribution with a mean of zero and precision of 0.0625 (corresponding to a SD of 4). With this parameterization of the prior distribution, over 98% of the viable parameter values fall between $-10$ and 10. Values at the extreme ends of such a prior distribution are highly unlikely to occur in logistic regression involving categorical and standardized continuous predictors and therefore represent a truly uninformative prior distribution. Priors for the subject and passage random intercepts in each study were set as a normal distribution with a mean of zero and precision of $\tau_s$ and $\tau_p$, respectively, where $\tau_s$ and $\tau_p$ were estimated from the study data and therefore had their own priors that were parameterized as a gamma distribution with shape and rate values set to 1. With this parameterization, the prior distributions for the random intercept precisions were constrained to have all positive values, with over 99% of values falling between 0 and 5.

Recall that in Study 2b, we employed an imputation modeling approach to address the issue of missing Versant™ scores. Due to the added complexity of the imputation models (described below), some tightening up of the prior distributions was required for the models to converge. These new priors were selected to constrain the range of plausible parameter values, while still allowing the data serve as the primary source of information driving the posterior distributions. For the intercept, the prior was set as a normal distribution with a mean of zero and a precision of 0.2 (SD of approximately 2.24), thus placing the vast majority of plausible values between $-5$ and 5. This prior specifies that there is still nearly complete coverage of the probability distribution on the log-odds scale, and therefore this is still a weakly informative prior distribution. For all beta parameters, the priors were set as a normal distribution with a mean of zero and a precision of 0.95 (SD of approximately 1.03). This specification is still only weakly informative, with the majority of viable parameter values falling between $-3$ and 3.
MCMC details

Here we provide technical details on the Markov chain Monte Carlo (MCMC) simulations of the posterior distribution. For each analysis, we simulated draws from the posterior distribution in three independent chains. We set the burn-in to 1000 samples (i.e., the initial 1000 draws were excluded from the final posterior distribution) to allow the MCMC algorithm to stabilize and achieve convergence. We set the thinning parameter to 50—that is, we only kept every 50th draw—to prevent autocorrelation. We kept 50 samples from each chain, leading to a final MCMC sample of 150 draws for each model parameter.

For each reported model, we confirmed that the chains were sufficiently converged by examining visual plots of the chains for each model parameter. In all cases, autocorrelations were low (nearly zero at a lag of 1) indicating that the sampled parameter values were independent draws from the posterior distribution. Visual inspection of the plots indicated that the chains were well-mixed, a necessary condition for convergence of the algorithm on the true posterior. This was confirmed by computing the $R$ value, which is the ratio of between-chain variance to within-chain variance. Convergence of the chains implies similar between- and within-chain variance (i.e., a $R$ value at or near 1.0), such that an $R$ value well above 1.0 indicates that the chains have not converged. For the reported analyses, all $R$ values were at or below 1.10, which has been suggested as a rule of thumb for indicating good mixing (Gelman, Carlin, Stern, & Rubin, 2004).

For Study 2b, the additional uncertainty and estimation imposed by the imputation modeling procedure required modifications to the MCMC sampling parameters to achieve a stable, converged posterior distribution. Most notably, the burn-in was increased from 1000 to 10000 in order to allow the MCMC algorithm to stabilize and converge. We also increased the number of chains from three to five and increased the number of kept samples from 50 to 150. Thus the final posterior for Study 2b contained 750 samples (vs. 150 in Studies 1 and 2a).

Model interpretation

In logistic regression, due to the nonlinear logistic curve function, model parameters are interpreted in log-odds space. Gelman and Hill (2007) suggest as an alternative to the log-odds interpretation of model parameters that the model parameter divided by 4 provides an estimate of the maximum possible change in probability correct given a one unit change in a predictor. For example, for Studies 2a and 2b, if the $\beta_1$ parameter were estimated to be 0.60, then following Gelman and Hill, we could interpret this parameter as indicating that switching from short, medium density passages (i.e., the baseline reference condition) to medium length, medium density passages would correspond to a maximum change of 15% accuracy (.60/4 = .15 probability correct). We used this interpretation to define the Region of Practical Equivalence (ROPE) as described below.

When examining the posterior distribution, the median value in the posterior distribution for a parameter is roughly equivalent to the maximum likelihood estimate for that parameter when using weakly informative priors, as we have done here. The 95% highest density interval (HDI) indicates the range of most credible values for the parameter, and therefore is much like a confidence interval. That is, similar to null hypothesis significance testing, if the HDI does not include zero, then we can conclude that the parameter is different from zero. As stated above, we can also set ROPEs to define a range of values that would be considered equivalent to a null effect in practice (Kruschke, 2011). If the HDI falls entirely within the ROPE, then we can conclude that the parameter is practically equivalent to zero.

To define the ROPE, we assumed that if a switch between two conditions (or, for working memory and Versant™ scores, a one $SD$ change in working memory ability or L2 listening ability, respectively) corresponded to a maximum change of $+/−$ 5% accuracy or less, this would be considered a minimal change and would be practically equivalent to a null effect. This corresponds to a beta parameter of 0.20 (since $.20/4 = .05$). Therefore, we set the ROPE at $+/−$ 0.20 for all beta parameters.
Imputation procedure

In Study 2b, approximately 15% of the sample was missing their Versant™ data. To address this, we employed an imputation modeling approach, by which we imputed a plausible value for any missing data point at each step in the MCMC sampling algorithm. This imputed value was then included in the substantive analysis (i.e., the logistic regression). When performing missing data imputation, it is best to incorporate into the imputation model any available variables that are related to the variable that is missing data. This allows the imputation model to take into account any known relationships between the imputation predictors and the missing variable when computing a plausible replacement value at every step in the imputation procedure.

Preliminary analyses and theoretical hypotheses identified the following variables as the most relevant for inclusion in the imputation model for the Versant™ scores: working memory, overall accuracy on the listening comprehension task (i.e., percent correct across all conditions), age, age when started learning the L2, level of education (binary variable, where 1 = more than high school education, 0 = no higher than high school education), and six of the questions from the ILR Self-Assessment of Foreign Language Listening Proficiency (after being translated into Spanish).13

To impute values of the Versant™ score within the Bayesian framework, for each MCMC sample, a given participant’s missing value was imputed by drawing from a normal distribution with mean, $\mu_{\text{Versant}}$, and precision, $\tau_{\text{Versant}}$, where $\mu_{\text{Versant}}$ was estimated by a regression equation involving the variables identified in the preliminary analysis, and $\tau_{\text{Versant}}$ was estimated from the data, with its own prior distribution set as the gamma distribution with shape of 4 and rate of .125, which corresponded to positive values falling primarily in the range of 0 to 50.

The variables of age, age when started learning the L2, high school education, and the ILR variables also contained a small amount of missing data themselves (minimum of 1 participant, maximum of 10 participants). Therefore, separate imputation models were similarly constructed for these variables. For the normally distributed variables (age, age when started learning the L2), the imputation model’s distribution was determined similarly to the Versant™ scores, with a mean determined by a regression involving all other predictors in the imputation models (including versant), and a precision estimated from the data, using the same priors for the precision as for the Versant™ imputation model. For binary variables (high school education and the six ILR variables), missing values were imputed by sampling from a Bernoulli distribution with a probability parameter estimated by a logistic regression involving all other imputation predictors.

Prior to fitting the final models, the effectiveness of the imputation modeling procedures were evaluated by comparing the results of the substantive regression model (i.e., the regression equation for Study 2b listed at the beginning of this appendix) when fitted to three subsets of the full dataset: (1) an “artificial complete dataset” containing only data from participants with observed Versant™ scores (n = 142); (2) an “artificial missing dataset” in which we artificially created missingness in the artificial complete dataset by removing data for 22 participants (roughly 15% of the artificial complete dataset) in order to match the proportion of missing data in the full dataset; and (3) an “artificial dropped-cases data” (n = 120) for which we excluded those 22 cases in the artificial missing dataset with missing Versant™ scores, to simulate the effect of dropping incomplete cases from the analysis. The results of the model fitted to the artificial complete dataset provide an estimate of the “true” relationships (i.e., $\beta$ coefficients) for this artificial dataset that contains no missing data. If the imputation modeling procedure is working as intended, the artificial missing results should look similar to the artificial complete results, and a comparison of these two results can provide insights into the impact of the additional uncertainty introduced in the missing-data analysis. Note that, for the artificial missing-data analysis, some differences from the artificial complete-data analysis are expected in the estimated parameters due to the fact that we are appropriately incorporating the uncertainty we have regarding the imputed data values. Finally, a comparison of the artificial dropped-cases results to the artificial complete results demonstrates how dropping cases with missing data would impact our inferences regarding the model parameters (i.e., effects).

13 One of the can-do statement was for ILR Level 1 (“basic directions and instructions”) and the remaining five were for ILR Level 3 (“follows all conversations between native speakers”, “rarely asks to repeat”, “inferential/indirect meaning”, “debates, proposals, and speculations”, and “opinion and points to support it”). See Appendix G.
Table H-1 reports the model parameters and HDIs for the simple effect of L2 listening ability (the variable with missing data) on task performance. There are two important observations to note. First, the imputation modeling approach produced much more accurate (i.e., valid) model parameters than the dropped-cases analysis, which underestimated the magnitude of the true relationship. Second, the HDI widths for both the artificial missing (with imputed data) and the artificial dropped-cases analyses are larger than the artificial complete-data analysis, and the HDIs are similarly wide for both approaches. This is expected, given that the imputation approach incorporates additional uncertainty regarding the imputed values, whereas the dropped-cases approach loses data – and therefore statistical power – by excluding cases.

### Table H-1. Model parameters ($\beta$s) and HDI widths for the relationship between Versant™ and accuracy, as estimated by the substantive model.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>$\beta$</th>
<th>Width of HDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial complete-data</td>
<td>0.383</td>
<td>0.024</td>
</tr>
<tr>
<td>Artificial missing-data with imputation</td>
<td>0.382</td>
<td>0.030</td>
</tr>
<tr>
<td>Artificial dropped-cases data</td>
<td>0.348</td>
<td>0.030</td>
</tr>
</tbody>
</table>

*Note. HDI = highest density interval.*

A second important source of evidence that the imputation modeling approach was functioning properly comes from an examination of the model parameters and HDI widths for the relationship between WM (which had no missing data) and accuracy. See Table H-2 for these results. Again, an examination of the model parameters and HDI widths indicates that the imputation modeling approach was functioning properly. First, as with the Versant™ scores, the imputed-data parameter estimates for the simple effect of WM were substantially more accurate than those from the dropped-cases analysis, which overestimated the magnitude of the WM—accuracy relationship. Second, the HDI widths for the imputed analysis are wider than the complete-data analysis; as expected, because WM and Versant™ scores are somewhat related, the additional uncertainty in the Versant™ scores was carried through to the substantive analysis and (appropriately) impacted our certainty in the WM—accuracy relationship. Note also that the dropped-cases HDIs were the widest of all three analyses, further suggesting that simply dropping cases with missing values Versant™ scores would have negatively impacted our estimate of the WM effects (likely due to the smaller sample size).

### Table H-2. Model parameters ($\beta$s) and HDI widths for the relationship between working memory and accuracy, as estimated by the substantive model.

<table>
<thead>
<tr>
<th>Analysis</th>
<th>$\beta$</th>
<th>Width of HDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial complete-data</td>
<td>0.522</td>
<td>0.160</td>
</tr>
<tr>
<td>Artificial missing-data with imputation</td>
<td>0.591</td>
<td>0.176</td>
</tr>
<tr>
<td>Artificial dropped-cases data</td>
<td>0.688</td>
<td>0.227</td>
</tr>
</tbody>
</table>

*Note. HDI = highest density interval.*

To summarize, in this artificial simulation of missing data using a subset of the real data, we incorporated as much information as available regarding WM, Versant™, and task performance within the analysis by imputing plausible values for the missing Versant™ scores, rather than simply excluding cases with missing data. This imputation modeling approach provided more valid estimates of the true relationships between task performance and both Versant™ (i.e., the variable with missing values) and WM (i.e., a variable with no missing values). In contrast, the standard practice of dropping cases with missing values not only introduced uncertainty to the resulting model parameters, but in fact it produced inflated estimates of the WM effects.

### Summary

We opted to use Bayesian analysis methods because they provided a coherent approach that both supported the goals of the study (e.g., estimating our confidence in a predicted null effect) and addressed peculiarities of the dataset (i.e., missing data). We set weakly informative priors to allow the data to be the primary determinant
of the inferences drawn from the analyses. A simulation test run demonstrated that the imputation modeling procedure adequately preserved the multivariate relationships between key variables in the analysis, while appropriately propagating and incorporating uncertainty through the entire analysis in a coherent, unified manner.
APPENDIX I: DEMOGRAPHIC DESCRIPTORS FOR ALL STUDIES

Study 1

Table I-1. Descriptive statistics Study 1 participants

<table>
<thead>
<tr>
<th>(n=52)</th>
<th>Age (n=51, 1 omitted)*</th>
<th>Length of residence in L2 country (yrs) (n=46, 6 omitted)*</th>
<th>Time spent in school/work environment where L2 is used (yrs) (n=43, 9 omitted)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>30.0 (8.3)</td>
<td>3.5 (4.8)</td>
<td>6.2(4.8)</td>
</tr>
<tr>
<td>Min</td>
<td>18.0</td>
<td>0.0</td>
<td>.08</td>
</tr>
<tr>
<td>Max</td>
<td>62.0</td>
<td>20.0</td>
<td>21</td>
</tr>
<tr>
<td>Median</td>
<td>28.0</td>
<td>1.0</td>
<td>4</td>
</tr>
</tbody>
</table>

*omitted participants did not provide this information

Table I-2. Highest level of education attained Study 1 participants

<table>
<thead>
<tr>
<th>(n=52)</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>High school</td>
<td>10</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Some college</td>
<td>5</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>Professional training</td>
<td>2</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td>College</td>
<td>23</td>
<td>45</td>
<td>78</td>
</tr>
<tr>
<td>Some graduate school</td>
<td>1</td>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>M.A.</td>
<td>8</td>
<td>16</td>
<td>96</td>
</tr>
<tr>
<td>Ph.D./M.D./J.D</td>
<td>1</td>
<td>12</td>
<td>98</td>
</tr>
<tr>
<td>(missing)</td>
<td>1</td>
<td>2</td>
<td>100.0</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

Study 2a

Table I-3. Descriptive statistics of Study 2a participants

<table>
<thead>
<tr>
<th>Age (n=172)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
</tr>
<tr>
<td>Min</td>
</tr>
<tr>
<td>Max</td>
</tr>
<tr>
<td>Median</td>
</tr>
</tbody>
</table>
Study 2b

Table I-4. Descriptive statistics of Study 2b participants

<table>
<thead>
<tr>
<th></th>
<th>Age (n=161, 8 omitted)*</th>
<th>Length of residence in L2 country (yrs) (n=144, 25 omitted)*</th>
<th>Time spent in school/work environment where L2 is used (yrs) (n=140, 29 omitted)*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>37.6 (11.4)</td>
<td>12.5 (10.1)**</td>
<td>7.5 (7.03) **</td>
</tr>
<tr>
<td>Median</td>
<td>38</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>S.E.</td>
<td>0.9</td>
<td>0.84</td>
<td>0.59</td>
</tr>
<tr>
<td>Min</td>
<td>18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max</td>
<td>66</td>
<td>59</td>
<td>38</td>
</tr>
</tbody>
</table>

*Omitted participants did not respond to questionnaire item

**Decimals represent months reported by participants

Table I-5. Highest level of education attained for Study 2b participants

<table>
<thead>
<tr>
<th></th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than High School</td>
<td>2</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>High School</td>
<td>33</td>
<td>0.2</td>
<td>0.21</td>
</tr>
<tr>
<td>Professional Training</td>
<td>16</td>
<td>0.09</td>
<td>0.3</td>
</tr>
<tr>
<td>Some college</td>
<td>36</td>
<td>0.21</td>
<td>0.51</td>
</tr>
<tr>
<td>College</td>
<td>47</td>
<td>0.28</td>
<td>0.79</td>
</tr>
<tr>
<td>Some Graduate School</td>
<td>5</td>
<td>0.03</td>
<td>0.82</td>
</tr>
<tr>
<td>Masters</td>
<td>12</td>
<td>0.07</td>
<td>0.89</td>
</tr>
<tr>
<td>Ph.D./M.D./J.D.</td>
<td>3</td>
<td>0.02</td>
<td>0.91</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>0.02</td>
<td>0.93</td>
</tr>
<tr>
<td>(missing)</td>
<td>11</td>
<td>0.07</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>169</td>
<td>1.00</td>
<td>1</td>
</tr>
</tbody>
</table>
Corresponding Author and Reprints:
Sarah Wayland, PhD, University of Maryland Center for Advanced Study of Language, (301) 226-8938, swayland@casl.umd.edu, www.casl.umd.edu.

Funding/Support: This material is based upon work supported, in whole or in part, with funding from the United States Government. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government. Nothing in this report is intended to be and shall not be treated or construed as an endorsement or recommendation by the University of Maryland, United States Government, or the authors of the product, process, or service that is the subject of this report. No one may use any information contained or based on this report in advertisements or promotional materials related to any company product, process, or service or in support of other commercial purposes. This report is not Releasable to the Defense Technical Information Center per DoD Directive 3200.12. The Contracting Officer’s Representative for this project is David Cox, Government Technical Director at CASL, (301) 226-8970, dcox@casl.umd.edu. The Technical Task Order Manager for this project is Susanne Whitt, PhD, Senior Analyst/Professor—ES/Research and Analysis Division, Defense Language Institute Foreign Language Center, (831) 393-9279, susan.a.whitt3.civ@mail.mil.

Acknowledgments: We wish to thank the following people from the DLIFLC: Donald C. Fischer, PhD, Provost; Shannon Salyer, PhD, Senior Research Scientist - Research and Analysis; John A. Lett Jr., PhD, Dean - Research and Analysis; James Dirgin, former Chief of Test Review and Education; Ruth Mehr, Chief of Test Development; Gerd Brendel Acting Chief of Test Review and Education.

We wish to thank the DUIELC for their incredible support in providing participants and a facility in which to test them. Sally Carter, Laura MacKenzie, Frank Lawrence, Larry McLeod, Anthony Sabbs, Phyllis Santleben, Miles Witt, and Wende Smith were particularly helpful to us, though there were many others who helped us accomplish our goals. The Liaison Officers for the Arabic countries were supportive as well; we couldn’t have done this without them. Last, but not least, we are grateful to the students at DUIELC who participated in our study.

We also wish to thank Catherine J. Doughty, PhD, Area Director for Second Language Acquisition at CASL, for her calm guidance, as well as her careful and thoughtful review of this document. We are grateful to CASL’s IT staff and especially Beth Lerie, Michael McGrath, Patrick Allen, and Jason White for their help with all things technical. We appreciate Alyssa Endres and all her hard work with testing of participants in San Antonio.