Technical Details

TTO 2012 2.6 ▪ CDRL A017 ▪ DID DI-MISC 80508B ▪ Contract No. H98230-07-D-0175

The effects of note taking on foreign language listening comprehension

An empirical study

Martyn Clark, PhD, Sarah Wayland, PhD, Peter Osthus, MA, Kassandra Gynther Brown, BA, Shaina Castle, BA, Alexandra Ralph, BA[1]

In an experimental study, CASL researchers determined the extent to which allowing note taking on a foreign language listening comprehension test, which mimics the Defense Language Proficiency Test (DLPT), affects test performance. Research participants listened to DLPT-like passages in both an *Allow Notes* and a *Listening Only* condition. If note taking is beneficial in such a situation, the results should show better performance in the *Allow Notes* condition than in the *Listening Only* condition. However, the results failed to show a difference between these two conditions. In addition, a closer examination of the data suggests that even those participants who felt that notes were helpful did not consistently perform better.

This report starts with a brief review of the literature looking at the role of note taking during listening in the native language and in a foreign language. It then describes the design of the experimental study. We used a within-subjects research design to minimize the effects of idiosyncratic participant profiles. An analysis of the data follows, and we conclude with a discussion of the implications of the results.

LITERATURE REVIEW

Note taking during native language listening

Most of the research that measures the impact of note taking while listening to native language (or L1) academic material (e.g., passages, lectures) indicates that having the notes available while answering test items is the most beneficial aspect of note taking, not the act of taking notes alone (Dunkel, Mishra, & Berliner, 1989; Hartley, 1983; Kiewra, 1985). Reviewing notes before taking a test can also be beneficial, even if the notes are not available during the test itself. In an early empirical study of note taking, Crawford (1925a) found that taking notes on a lecture and then reviewing those notes prior to taking a quiz had a positive correlation with

[1] Authors are listed by degree (PhD, MA, BA), and alphabetically within degree. Contributions to the project are as follows: Martyn Clark, PhD (principal investigator); Sarah Wayland, PhD (materials development, experimental design, writing); Peter Osthus, MA (data analysis, graphics generation, data collection, writing); Kassandra Gynther Brown, BA (materials development, data collection, editing, and administrative support); Shaina Castle, BA (literature review, test design, materials development, data management, technical editing, and administrative support) Alexandra Ralph, BA (data management, editing, administrative support).

quiz results when the notes taken contained information that was needed to perform well on the quiz; essentially, "…the quiz results of these students were very definitely limited to the points upon which they took notes (p. 289)." For native speakers listening to their L1 in an academic setting, note taking can be beneficial as a repository of information for later use. In a series of follow-up studies with various note taking and testing conditions, Crawford (1925b) found additional support for the benefits of note taking for remembering information. A more recent summary of note taking research (Hartley, 1983) also reported positive effects of reviewing previously taken notes on information recall in all but six of the 28 studies reviewed, with the other studies showing no effect.

Though note taking is useful for encoding information for later use, this does not mean that note taking in the L1 is universally beneficial. Even Crawford's early investigations into note taking revealed a detrimental effect in some conditions, namely on immediate recall as measured by a true/false test when compared to a no notes condition (Crawford, 1925b). In the note taking research summary previously cited, Hartley (1983) reported that, of the 57 studies examining the impact of note taking (without review) on recall, 19 showed no effect of note taking on recall, and 4 studies showed a detrimental effect. In a similar review of 22 note taking studies in non-academic contexts – jury situations, counseling, interviewing – Hartley (2002) also found that while results were generally positive (14 out of 22), some studies reported no clear effects (5 out of 22) or even negative effects (3 out of 22) for note taking. Thus, though note taking can be useful for capturing information for later review and use, the benefits of note taking are mitigated if the act of note taking detracts from the initial processing of the information, even for native speakers taking notes in their L1.

Note taking during second language listening

Research on note taking during second language (or L2) listening also shows inconsistent effects. Most of the research examining the effects of note taking while listening to L2 materials has found that note taking has no impact on overall performance, but some research indicates that note taking can be beneficial for some types of passages or under certain circumstances.

Language proficiency

In a comparison of note taking between native and non-native speakers of English, Dunkel et al. (1989) assigned native and non-native participants to a listening only or listening and note taking condition and asked them to listen to a 22-minute lecture. Notes were collected immediately following the lecture. Results of a subsequent multiple-choice test on the contents of the lecture showed that native speakers of English outperformed non-native speakers of English on recalling English lecture concepts and details regardless of whether the native-English speakers took notes or not. Unfortunately, Dunkel et al. (1989) did not use a language proficiency measure to indicate the differences in English proficiency level of the non-native speakers, so a more refined conclusion cannot be made about note taking's impact on recall for differing levels of L2 proficiency.

One research study suggests that access to previously taken notes does not benefit learners whose TOEFL score is high enough to gain admission to an English-medium university. Using intact university-level English as a Second Language (ESL) classes, Chaudron, Cook, and Loschky (1988) found that the availability of previously taken notes had no impact on L2 English students' performance on multiple choice tests that were taken immediately after listening to 6-7 minute L2 lectures. In the study, 98 non-native-English speaking students were provided note taking paper while they listened to a pre-recorded English lecture on an academic topic. Half of them were randomly selected to be allowed to keep their notes for review as they completed 20 multiple-choice questions and a cloze comprehension test. All students did this three times, listening to a different lecture and being randomly assigned to a test-with-notes or test-without-notes condition each time. The researchers suggest that students who were not allowed to keep their notes could still rely on their immediate recall to answer the test questions. However, the study did not include a no-note taking condition, so it is impossible to determine from this work whether the act of taking notes helped or hindered the participants' initial comprehension.

In a study which examined the effect of note taking on L2 listening comprehension performance and its interaction with lower and higher language proficiency (based on a median split on a paper-and-pencil

Institutional TOEFL listening section), Carrell, Dunkel, and Mollaun (2002) found a significant effect for proficiency, with participants above the median proficiency scoring statistically significantly higher on a computer-based listening test than those below the median. There was also a significant interaction between proficiency and lecture length. However, there were no interaction effects between proficiency and note taking condition (listening only, listening and note taking). In their discussion, the authors suggest that there may be a proficiency floor below which the effects of note taking will not be observed.

Cognitive effort

Because writing operates at a fraction of the speed of speaking (writing speed is 0.2 to 0.3 words/second and speaking is 2 to 3 words/second), note taking introduces considerable time-pressure (Piolat, Olive, & Kellogg, 2005). In one study, participants took notes faster (as measured by reaction times) in their L1 (French) than their L2 (English) when listening to English passages (Barbier & Piolat, 2005; described in Piolat et al., 2005). This suggests that taking notes while listening to an L2 passage requires more cognitive effort than taking notes on a passage presented in the L1 due to the additional cognitive demands of working in the L2. However, although Dunkel et al. (1989) did report a significant effect for working memory capacity on information recall, they found no significant interaction between working memory capacity and note taking. In a study of note taking with learners of English as a Foreign Language (EFL) in Taiwan, Lin (2006) also found working memory capacity to be a significant predictor of overall listening comprehension, while note taking was not. However, for passages that are only heard once, there may be some benefit to note taking, as note taking allows the listener to capture the ephemeral bits of information in the passage (Chaudron, Loschky, & Cook, 1994; Lin, 2006), and L2 listeners can off-load information from the passage into their notes rather than having to mentally retain all information from a single presentation.

Passage speech rate

In her dissertation study on the effect of note taking and speech rate on L2 comprehension, Lin (2006) examined the impact of note taking on the comprehension of L2 passages that were presented at faster (180 wpm) and slower (120 wpm) speech rates. Participants listened to English passages delivered with fast or slow speech rates in note taking and no-note taking conditions. Though note taking did not significantly affect overall performance on multiple choice comprehension items, separate regression analyses for each speech rate did find that note taking was a statistically significant predictor for comprehension of the faster speech rate but not the slower one. This result is somewhat hard to interpret, as some students in the note taking condition took fewer notes or even no notes for the faster speech rate passages.

Passage length

Intuitively, one would expect that listeners would benefit more from note taking on longer passages than shorter passages, as information could be preserved in the notes. However, Carrell et al. (2002) found that this was not the case; for both higher and lower (above and below the median split on the Institutional TOEFL) proficiency participants, note taking did not significantly impact performance on long (~ 5 minutes) passages. Conversely, a positive impact for note taking was found for shorter (~ 2.5 minutes) passages for both proficiency groups. It is important to note that though this result was statistically significant, the effect size of this result was rather small.

Passage topic

Though it was not the main research thrust of the study, Carrell et al. (2002) found an effect for note taking on performance when comparing across different passage topics. On a computer-based test that was similar in design to the computer-based TOEFL, Carrell et al. (2002) found that note taking was beneficial for arts and humanities passages, but not for physical science passages. This effect was also mediated by passage length, as indicated above. The authors noted that this effect may be at least partially attributable to a larger portion of their participant population being scientifically-oriented (e.g., intending to major in the sciences) rather than humanities-oriented.

Forced note taking

The studies reviewed thus far have generally tried to determine if note taking has a positive impact on performance. A non-significant finding for the note taking condition is interpreted as note taking having no effect. However, Hale and Courtney (1994) found that insisting that participants take notes while listening to L2 material negatively impacted their performance. While listening to multiple L2 passages, each participant took notes for half of them. Some participants were merely permitted to take notes, and some participants were urged to do so. Note taking had no impact on performance on comprehension items when it was simply allowed; but, when note taking was urged, performance was worse than when no notes were taken at all. In spite of this, participants reported feeling more at ease when they were able to take notes than when they were not, and they believed taking notes helped them remember more information.

Content and quality of the notes

Other studies have examined the impact of the content and quality of notes of L2 listeners on listening performance beyond the mere act of taking and/or reviewing notes (Carrell, 2007; Chaudron et al., 1988; Clerehan, 1995; Dunkel, 1988; Hayati & Jalilifar, 2009; Song, 2012; Tsai & Wu, 2010). As might be expected, students who (a) took "better" notes (Dunkel, 1988; Song, 2012), (b) took notes that related to material that was subsequently tested (Carrell, 2007; Chaudron et al., 1988; Dunkel, 1988), or (c) took notes using a specific note taking method (Hayati & Jalilifar, 2009; Tsai & Wu, 2010) tended to perform better. Unfortunately, as many of these studies did not include a no-note taking condition, it is impossible to determine if taking "good" notes created a performance advantage over taking no notes. Further, Carrell (2007) posited that "[i]t may be unrealistic to expect any but the most advanced L2 learners to produce quality notes …" (p. 45), and Song (2012) also commented that a test taker's quality of notes may reflect the test taker's L2 listening proficiency.

Summary

Overall, the literature suggests that the ability to take notes can occasionally be advantageous for L2 listening comprehension under certain circumstances, though there are numerous additional factors that can affect this relationship (e.g., Bloomfield at al., 2010). The findings of Lin (2006) and Hale and Courtney (1994) suggest that note taking strategies affect individual L2 listeners differently. Many studies indicate that having notes on hand during an L1 test is beneficial (Hartley, 1983; 2002), especially when the content of the notes matches the items being tested. One would expect this to generally hold true in the L2 as well. Though Chaudron et al. (1988) did not find an effect for having notes on hand, they did indicate that items were more likely to be answered correctly when notes were available and contained the relevant information. At worst, there was no effect of note taking on test performance; the effect also may vary depending on the type of passages presented (Carrell et al., 2002). The limited research exploring the effects of optional note taking on L2 listening comprehension suggests that simply permitting note taking is not detrimental to the examinees, but examinees who feel forced to take notes may not perform as well (Hale & Courtney, 1994).

Hypotheses

There does not seem to be a clear-cut prediction for the potential effect of note taking on a DLPT-like listening test. The DLPT uses relatively short, authentic passages, each played twice, and examinees have access to the questions before listening to the passages. In the note taking literature, most studies did not use truly authentic materials. One study that did use authentic materials presented a much longer passage than those found on the DLPT (Dunkel et al., 1989). Also, none of the research examined the impact of note taking on performance when test questions are provided before the listening passage. A study of the effect of note taking under DLPT-like testing conditions – relatively short, repeated passages with access to test questions while listening – is needed to more clearly observe, analyze, and predict the effect of note taking in such a context. Based on the mixed results from the literature review above, we designed an experiment that would examine note taking in the DLPT-context. This enables us to explore two questions:

- Does allowing note taking influence performance on a DLPT-like L2 listening comprehension test?
- Do participants who actually take notes do better when allowed to take them?

DESIGN AND PROCEDURE

A hallmark of the DLPT listening subtest is that examinees listen to passages in the target language and answer test questions in English, which is the L1 for most examinees. To reproduce this relationship for this study, English was chosen as L2, with L1 Spanish speakers providing the examinee population. Thus, the materials described in this report were designed to measure listening proficiency in English as a second language.

Participants

All data were collected in Austin, Texas. Participants were recruited through advertisements within the local Spanish-speaking communities. A participant recruiting firm was contracted to identify potential participants, conduct pre-screener questionnaires for study eligibility, and schedule eligible participants. Eighty-nine participants were determined to be eligible and completed some or all of the study. Participants who failed to perform at chance level (25% accuracy) on the multiple-choice component of the listening test as well as those with incomplete data were removed. This yielded a total of 62 participants (35 female, 25 male, 2 unknown) with complete data on the listening test, the Versant™ Pro Speaking test, and the working memory measures (Blockspan and Shapebuilder).

Materials

Test materials were designed to fulfill two complementary aims: 1) to provide a reasonable facsimile of the type of multiple-choice items on the DLPT listening subtest in order to investigate note taking on a DLPT-like assessment without jeopardizing the security of the operational test, and 2) to capture additional examinee information during the testing process, namely information recall and working memory capacity. While this information is not part of the DLPT, it might help explain some facets of performance on the items. Many of the materials to be used in this study were originally developed for an investigation of factors affecting second language listening comprehension (Blodgett, Bloomfield, Wayland, O'Connell, Linck, Gynther, & Kramasz, 2011), and the following descriptions borrow from that foundational work.

Passages

The passages used in this study targeted proficiency at ILR listening levels of 2 and 3. These passages were primarily extracted from the National Foreign Language Center's (NFLC's) *Introduction to Passage Rating* course (2005), which provided passage rating levels for each passage based on the passage rating methodology developed by Child (1998). Two additional passages were taken from WTOP radio news broadcasts. Those passages were leveled internally by CASL researchers, and the leveling was later confirmed by a DLIFLC passage rating expert. All of the passages–originally intended for native speakers of English–represent a range of Final Learning Objective (FLO) topics. Table 1 presents information about the final versions of the passages after manipulations (described in Appendix A).
Each participant heard a total of 12 test passages, six at ILR level 2 and six at ILR level 3 (See Appendix B for a sample). Each passage was accompanied by two multiple-choice and four or more recall questions. The

multiple-choice questions were intended to mimic test questions on the DLPT. The recall questions were included to provide additional information about the participants' comprehension of the passage.

Table 1. Characteristics of Test Passages

Level	Short Title	# of Speakers	FLO Topic	Syllable Count	Word Count	Density*	Length
2	Refugees	2	Cultural and Social	231	131	0.73	1:04
2	Vitamin D	2	Scientific and Technological	221	146	0.72	1:06
2	Coyote	3	Geography: Physical, Political, Economic	225	146	0.70	1:13
2	AM/PM	2	Geography: Physical, Political, Economic	236	160	0.73	1:15
2	Taser	2	Cultural and Social	218	139	0.71	0:54
2	Traffic	2	Cultural and Social	226	156	0.72	1:03
3	Bill Cosby	3	Cultural and Social	355	228	0.73	1:34
3	Charter Schools	1	Cultural and Social	356	237	0.70	1:34
3	Secretary Rice	3	Military-Security	363	270	0.71	1:38
3	Kurds	2	Economic-Political	356	246	0.71	1:40
3	Terrorists	1	Military-Security	361	233	0.72	1:40
3	Artful Brain	1	Cultural and Social	355	243	0.70	1:34

*Type/token ratio as measured by MATTR (see Appendix A for description)

Comprehension Questions

For each passage, two or three multiple-choice comprehension questions were developed by a CASL researcher who had significant item writing experience. Every effort was made to match the test development guidelines in the "DLPT5 Lower Range Multiple-Choice Test Specifications" document, and to produce items that matched the item type distribution (global versus non-global) and skill assessed (e.g., main ideas, major details, implications). All items were developed to be appropriate given the ILR level of the passage. Items were circulated among CASL researchers for internal feedback and review, and revisions were made accordingly. Items were also given to several native speakers to see if they could be keyed without reference to the passage; problematic items were changed. Two items per passage were chosen for the final version of the materials. Example multiple-choice questions are shown in Appendix B. The finalized items were translated into Spanish (described below).

Recall Questions

Although recall questions are not used on the operational DLPT, we have included them in this study to provide additional information about the effect of note taking on listening. For this study, we chose a verbatim recall task in which examinees must produce a target word from the surrounding context, similar to a cloze test. The text is taken directly from the transcription of the passage, with a single target word deleted. An example item is shown in Appendix B. These items were reviewed by CASL staff and piloted by five native speakers of

English to determine if the target word could be guessed without listening to the passages, and problematic target words were replaced as necessary.

Materials Translation

Because this study uses English as the target language with an L1 Spanish population of examinees, it was necessary to translate the multiple-choice test items and test instructions into Spanish. The recall questions were not translated, as they require the surrounding English context. The translation was done through a professional translation vendor with experience working with testing materials. CASL researchers with knowledge of Spanish oversaw the translation process and provided feedback to the vendor.
After translation, the Spanish language versions of the 24 multiple-choice test items were given to six native Spanish speakers to determine if any of the items could be keyed without reference to the listening passages. For 22 of the items, there were three or fewer keys. For the other two items, five out of six of the native speakers chose the key without hearing the listening passage. Though one of these items had been revised after internal piloting, neither had been problematic in revised form. One of these items was revised to remove wording from the key that was also in the stem.

Language Skills Assessment

Participants were required to have English listening proficiency in the range of ILR 1+ to ILR 2+/3, with the majority of the participants around ILR 2/2+.[2]

Versant™ Pro Speaking test

The Versant™ Pro Speaking test was administered to obtain an objective measure of participants' English language proficiency. All instructions for the Versant™ Pro test were translated into Spanish to ensure comprehension of the tasks for participants.[3]

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 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 non-participatory. Indeed, every task in the Versant™ Pro is machine-scored and administered over the telephone. All participants 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.

The Listening Comprehension subscore was used as an objective measure of English listening comprehension level for the current study. The Versant™ Pro Speaking test scores were correlated with ILR level using the Common European Framework of Reference (CEFR). The CEFR aligns Versant English overall scores with CEFR levels on the CEFR scale of the Oral Interaction Skills with reasonable accuracy (Bernstein &

---

[2] A Level 2 in listening is also described as Limited Working Proficiency, while a Level 2+ is described as Limited Working Proficiency - Plus ("Interagency Language Roundtable Language Skill Level Descriptions: Listening," 1985).
[3] 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.

De Jong, 2001). A table published by the American University Center of Provence provides CEFR equivalencies to ILR levels. According to these criteria, people with Versant scores between 47 and 68 were likely to listen at an ILR level of 2. We used a slightly wider range of 41 to 68 on the Versant Listening Comprehension subscore based on Wayland et al.'s (2012) protocol. They used 16 DLIELC participants whose instructors had judged them as being at ILR 2 and had been assessed as ILR 2 in their Oral Proficiency Interview (OPI) to supplement the CEFR-based score correspondence. Based on these participants, Wayland et al. (2012) determined that a Versant Listening Comprehension subscore between 41 and 68 would correspond with an ILR Listening Level of 2. Our study used the same criteria.

Language background questionnaire

To help understand the language backgrounds of our participants, we asked them to complete a modified version of a language background questionnaire, the LEAP-Q (Marian, Blumenfeld, & Kaushanskaya, 2007), in Spanish (See Appendix C for the English version).

Post-test questionnaire

At the end of the experimental session, participants took a note taking survey similar to that used by Carrell, Dunkel, and Mollaun (2002). The survey consisted of 14 questions and was intended to provide information on the participants' previous experience with note taking as well as their impression of the helpfulness of note taking in answering test items (See Appendix D for English version).

Working memory capacity assessment

Two measures of working memory (WM) were included in the current study to investigate how WM factors related to overall performance on the comprehension items and how they mediated the effect of note taking on performance. Understanding a foreign language taps several general cognitive abilities. One important cognitive ability is working memory (WM), typically defined as the capacity to attend to, temporarily store, and process incoming information. Listeners with a greater WM capacity understand more of what they hear when they are listening to their non-native language than listeners with a lower WM 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 WM and L2 listening comprehension. The study revealed both a direct and an indirect impact of WM on syntactic processing: participants with higher WM 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. McDonald (2006) found that performance on a WM measure correlated significantly with the accuracy of grammaticality judgments of spoken L2 sentences. In sum, the available research evidence suggests that individuals' WM capacity affects L2 listening comprehension.

In this study, working memory capacity was assessed using two spatial tasks. In spatial WM tasks, participants are asked to simultaneously process and store spatial information. The use of two spatial WM tasks helped us minimize the influence of task effects. An important benefit of Blockspan and Shapebuilder is that they can be administered to speakers of any language. Instructions for both tasks were translated into Spanish.

Blockspan task description

The first WM task participants responded to was Blockspan (Atkins, Harbison, Bunting, & Dougherty 2012). In this task, participants are shown a 4 x 4 series of squares and are asked to remember the serial order in which a sequence of yellow blocks appeared on the grid (see Figure 1 below). Each block within a sequence flashes for one second, one at a time, in one of the cells on the 4 x 4 grid. Trials are segmented into sets by the appearance of a black square mask that covered the entire grid for one second. After viewing a series of locations flash in a given trial, participants are asked to recall the locations that the squares were flashed in the correct order by clicking the squares in the same order that they appeared. Participants complete 16 trials of length 2–20. The task increases in difficulty by increasing the trial length and by increasing the number of sets within each trial. For instance, for the first trial, there is one set with two stimuli in the trial. For the next trial, there is one set with three stimuli, the next trial has one set of four stimuli, and the next trial has one set with five stimuli. After this, trials are made more difficult by including two sets of two, three, four, or five stimuli. Then, trials are made more difficult by including three sets of two, three, four, or five stimuli. Finally, for the last 4 trials participants view four sets of two, three, four, or five stimuli. The dependent variable for this task was participants' score, which was computed as follows. Participants received 10 points for the first item correctly recalled, 20 for the second item in a row correctly recalled, 30 for the third item in a row correctly recalled, and so on (each additional item in a series correctly recalled given that previous items were 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 item in the sequence correctly recalled.

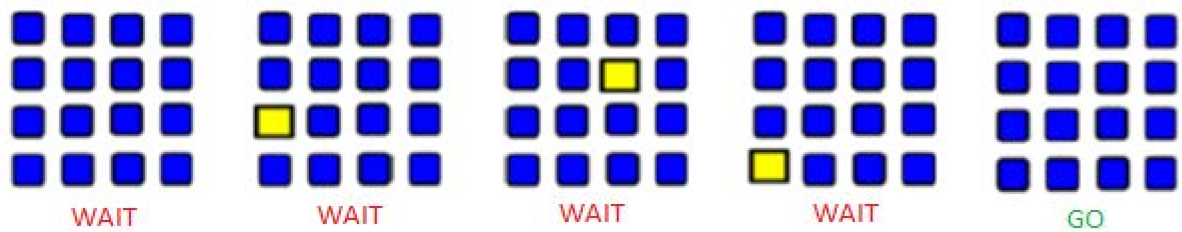

Figure 1. The Blockspan task

Shapebuilder task description

The second WM task was the Shapebuilder Task (Sprenger et al., 2013). This task is similar to Blockspan but involves the tracking of additional information. Participants are asked to remember the order and spatial position in which a series of colored shapes were presented. They see a 4 x 4 grid (see Figure 2) 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 are asked to remember the location, shape, and color of each item, and the order that items appeared. After the final item of a trial is presented, participants are asked to recreate the sequence by clicking on the correct colored shape and dragging it to the appropriate location. The Shapebuilder task increases in difficulty in two ways. First, trial length begins at two and increases to three and then four. Second, within each set of trials of a given trial length, the trials become more difficult by including more diverse stimuli of different colors/shapes. At the easiest level, items are all the same shape or color, and at the most difficult level, items are all different colors and shapes. Participants see the points awarded for each item immediately after releasing the mouse button.

Points are awarded as follows: participants receive 15 points for the first item correctly recalled and additional 15 points for every consecutive item correctly recalled in the sequence. Shapebuilder also awards points for partially recalled items; partial credit is only awarded when the correct location was guessed, such that the participant earns 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 is missed, the scoring starts 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 $r > 0.89$ (Sprenger et al., 2013). Furthermore, in Sprenger et al.'s work, Shapebuilder scores correlated with two previously validated measures of WM: Reading Span (Daneman & Carpenter, 1980) and Letter-Number Sequencing (Gold, Carpenter, Randolph, Goldberg, & Weinberger, 1997; Myerson et al., 2003), as well as with a measure of visuo-spatial WM called Blockspan (Atkins et al., 2012).

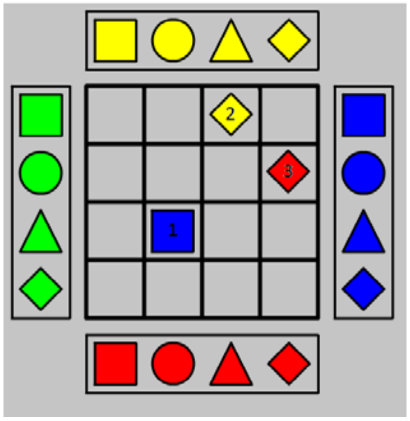

Figure 2. The ShapeBuilder task (Sprenger et al., 2013). The task is to remember the order in which a series of colored shapes have been presented. In this depiction, we indicate the order with numbers; the actual task presents the shapes one at a time, without the number.

Design

Independent and dependent variables

The independent variable for this study was the note taking condition, with two levels (*Allow Notes*/*Listening Only*). The dependent variables were the responses to the multiple-choice and recall questions on the listening test. Multiple-choice responses were treated as binary data (*correct* → 1, *incorrect* → 0) as were responses to the recall questions (exactly correct or semantically related → 1, semantically unrelated → 0). We also evaluated working memory capacity as a covariate.

Experimental conditions

As noted above, this research was conducted using a within-subjects design so that all participants experienced both the *Allow Notes* and the *Listening Only* conditions (See Table 1). The *Allow Notes* condition was the experimental condition, and the *Listening Only* condition was the control condition. Using a within-subjects design ensured that results from either individual condition will not be unduly influenced by the specific characteristics of any particular participant. Though carry over effects – the "bleeding over" of the experimental condition into the control condition – can be a concern for within-subjects designs (Levin, 1999), we do not believe that they represented any real threat to this particular study given the nature of our conditions. To control for order effects, some participants experienced the *Listening Only* condition first and the *Allow Notes* condition second, while other experienced the conditions in the opposite order. A schematic representation of the study is shown in Figure 3.

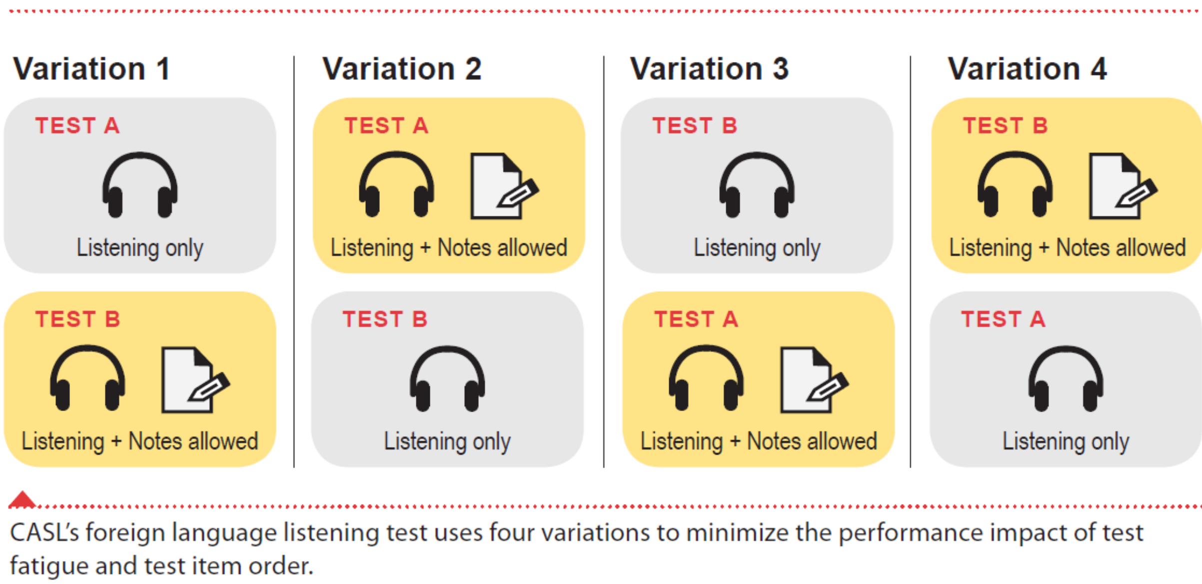

Figure 3. Test Delivery Design (pictorial design)

Procedure

Data were collected in groups consisting of one to eight participants. Upon being greeted by the researchers, participants were oriented to the nearest restrooms and then taken to the testing location. All participants were then given two consent forms, one to fill out and hand in to the researchers, and another for their personal records. All participants had been asked to complete the Versant™ Pro Speaking test prior to onsite testing; however, not all participants did so. Therefore, several participants completed their Versant™ test onsite. Because only one individual was able to take the Versant™ test at a time at the testing location, participants were interleaved throughout a given session to ensure that a given data collection session lasted no more than four hours in total duration. If a participant did not need to take the Versant™ test at the testing location, s/he first completed the modified version of the LEAP-Q. After completing the LEAP-Q, participants began the experimental portion of the study. Depending upon the test version, the participant either started with the *Allow Notes* condition or the *Listening Only* condition. After completing the first half of the listening test, participants were given the opportunity to take a 15-minute break before continuing on to the second half of the test. Directly following the listening test, participants completed the note taking questionnaire. Finally, participants completed the working memory tests, beginning with Blockspan and ending with Shapebuilder. Upon completion of the Shapebuilder task, participants were paid and debriefed on the purpose of the study. All study volunteers were paid $95 USD.[4]

RESULTS AND DISCUSSION

Demographic information

Our 62 participants' ages ranged from 19 to 65 years, with an average age of 37.2 years old. The average length of residence in a country where English was spoken was 12.5 years, and participants had spent an

[4] To ensure that participants took the test seriously, a slight deception was used in which participants were promised $70 for participation with an additional $1 for each correct answer on the multiple choice section. Participants were informed of this deception during the debriefing phase; all participants received the same payment upon completion of the tasks.

average of 7.3 years in a school or work environment where English was used. Of the participants tested, all but 3% had at least a high school diploma, while 54% had taken some post-secondary courses or had earned post-secondary degrees. See Tables E.1 and E.2 in Appendix E for more detailed demographic information.

Listening proficiency (English)

All participants were native speakers of Spanish with some non-native proficiency in speaking and listening to English. They were pre-screened to verify that they had at least an intermediate level of English language proficiency; they also took the Versant™ test during the experiment. The mean Versant™ Listening subtest score for the participants was 49 (out of 80), and the mean Overall Versant™ score was 49.29 (out of 80).[5] There were 13 speakers at ILR level 1, 46 at ILR Level 2, and 3 at ILR Level 3. Figure 4 shows the distribution of participants across the three ILR levels.

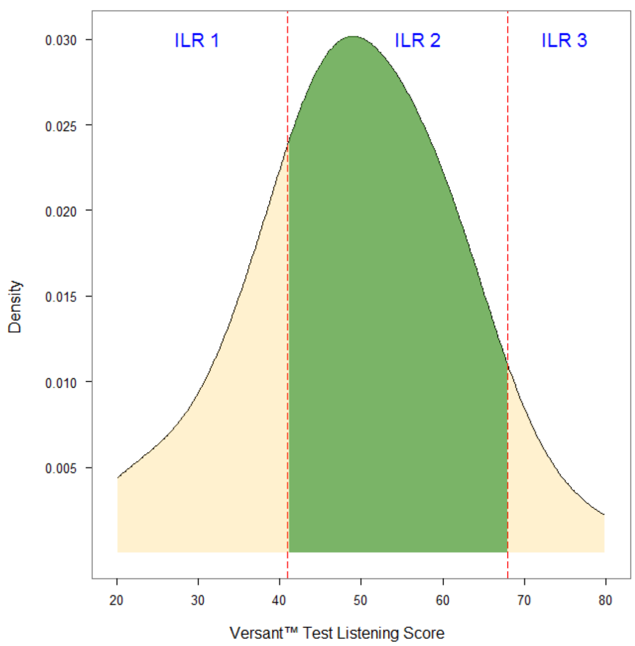

Figure 4. Density distribution of Versant™ test listening scores.

Effect of note taking

Below we present the data for the multiple choice items and the recall items. It is important to remember that performance on the multiple choice task is closest comparison to what performance would be on the DLPT. We included the recall data because previous work (Wayland, et al., 2012) has shown that it is more sensitive to manipulations than multiple choice measures.

Multiple-choice

Table 2 presents the average number correct on the multiple choice items for the participants out of a total of 24 possible.

[5] Because we are interested in listening proficiency, only the listening subtest score was used in the analyses. That the overall proficiency score was lower suggests that the participants' listening ability was higher than their productive ability.

Table 2. The descriptive statistics for the number (percentage) correct on multiple-choice questions (24 possible).

N	**Mean**	***SD***	**Minimum**	**Maximum**
62	14.7 (61.1%)	3.8 (16.0%)	7 (29.2%)	23 (95.8%)

Table 3 provides a summary of the mean percent correct across participants.

Table 3. Mean number (percentage) correct and standard deviation (SD) on multiple-choice questions as a function of notes condition (*Allow notes*, *Listening only*; 12 possible*)*

Length	**Mean**	**SD**
Listening only	7.4 (61.7%)	2.3 (19.1%)
Allow notes	7.3 (60.5%)	2.4 (20.2%)

Figure 5 shows the multiple-choice scores by condition (*Listening Only* vs. *Allow Notes*). Individual scores are shown as points ("jittered" to make them all visible); the means, and quartiles are shown using box and whisker plots. The mean is the horizontal bar inside the colored box. The upper and lower boundaries of the colored boxes correspond to the first and third quartiles (the 25th and 75th percentiles). The upper whisker extends from the third quartile to the highest value that is within one and a half times the distance between the first and third quartiles. The lower whisker extends from the first quartile to the lowest value within one and a half times the distance between the first and third quartiles.

The mean score for the *Listening Only* condition was very slightly higher than that of the *Allow Notes* condition. It is important to remember, however, that these boxplots do not take into account the potential contributing effects of working memory and listener proficiency.

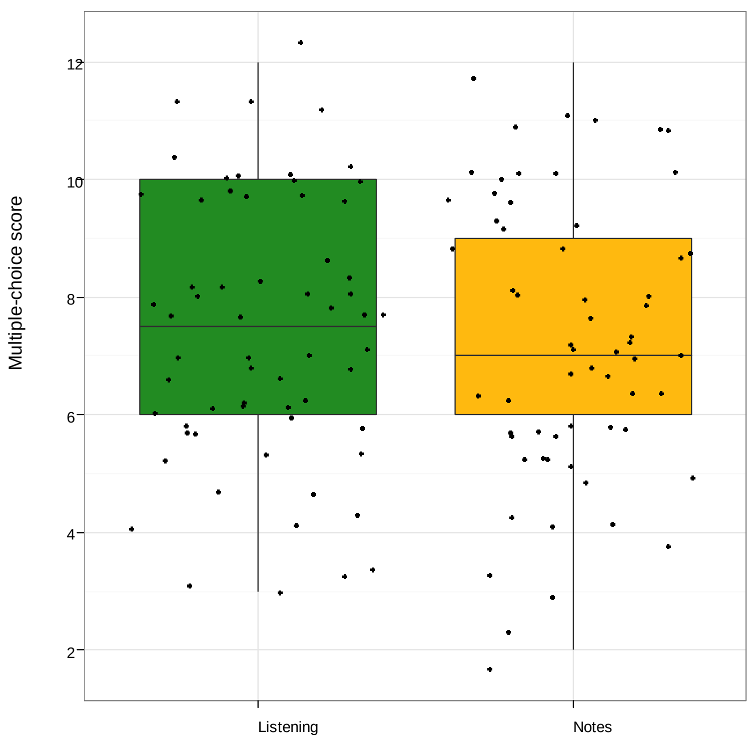

Figure 5. Multiple-choice scores by condition for all participants.

To determine whether the impact of allowing note taking changed when working memory capacity and listening proficiency were taken into account, we analyzed the data using a mixed-logit model (e.g., Bryk & Raudenbush, 1992; De Boeck et al., 2011; Doran et al., 2007; Jaeger, 2008). This approach allows for the modeling of repeated binary outcomes (correct/incorrect on test items) while appropriately accounting for nested, or partially crossed, factors, such as our *Allow Notes* condition.

Mixed-logit model

A random effects structure consistent with the study design and hypotheses was used to build the mixed-logit model. The model structure was selected in order to provide a direct test of the principal hypothesis concerning the study: Does allowing note taking influence performance on a DLPT-like L2 listening comprehension test? Multiple-choice question accuracy (Result) was modeled with note taking condition (*Allow Notes* vs. *Listening Only*), working memory composite score (WM), and Versant™ test listening score as fixed effects factors, and subjects and items as random effects factors. The model estimated the impact of these fixed effects factors while allowing random intercepts and random slopes for both random effects factors (subjects and items). All analyses were conducted within R, a statistical programming language (R Core Team, 2013), using the lme4 package (Bates, Maechler, & Bolker, 2012). The model as specified within the R computing environment is shown below, and the model results are provided in Table 4.

```
glmer(Result ~ Condition + WM + Versant Listening + (1+Condition|Subject) + (1+Condition|Item))
```
[6]

Table 4. Summary of the fixed effects in the mixed-logit model (Number of observations = 1488, log-likelihood = -856.7).

Fixed Effects	**Estimate**[1]	**SE**	**z value**	**Pr(>|z|)**	**Odds**
(Intercept)	-0.14	0.42	-0.33	0.74	0.87
Note taking Condition (Allow Notes vs. Listening Only)	-0.13	0.16	-0.80	0.42	0.76
Working Memory	0.29	0.05	6.39	0.00*	1.16
Versant™ Test Listening Score	0.02	0.01	2.60	0.01**	0.89

[1]The raw coefficients are expressed in logits (i.e., log-odds)
*Significant at *p*-value less than 0.01.
** Significant at *p*-value less than 0.001

As shown in Table 4 in the Pr(>|z|) column, there is no evidence that the ability to take notes during a DLPT-like listening test offers an advantage to the test taker ($p > .05$).[7] There is, however, evidence that performance changes depending on the listener's working memory capacity ($p < .01$) and proficiency ($p = .01$).

Impact of working memory

To ease interpretation of the model parameters, we converted raw scores on the Blockspan and Shapebuilder tasks to z-scores, and then combined the two scores to create a single working memory z-score for each participant. Figure 6 shows performance on the multiple choice assessments as a function of working memory and note taking condition (*Listening Only* vs. *Allow Notes*). Working memory clearly has an enormous impact on performance such that the accuracy of a participant's performance on multiple choice increases as his/her working memory increases. Furthermore, while there is some indication that taking notes is more advantageous to people with greater working memory capacity than those with less capacity, the high degree of overlap between the scores suggests that the differences appear minimal at best.

---

[6] We also analyzed the data using an interactive model (`glmer(Result ~ Condition * WM * Versant Listening + (1+Condition|Subject) + (1+Condition|Item)`). This model did not show a significant improvement in model fit over the simple main effects model described in the text.

[7] It is always difficult to "prove" a null effect. See Appendix F for Bayesian estimate of the mean difference between the conditions.

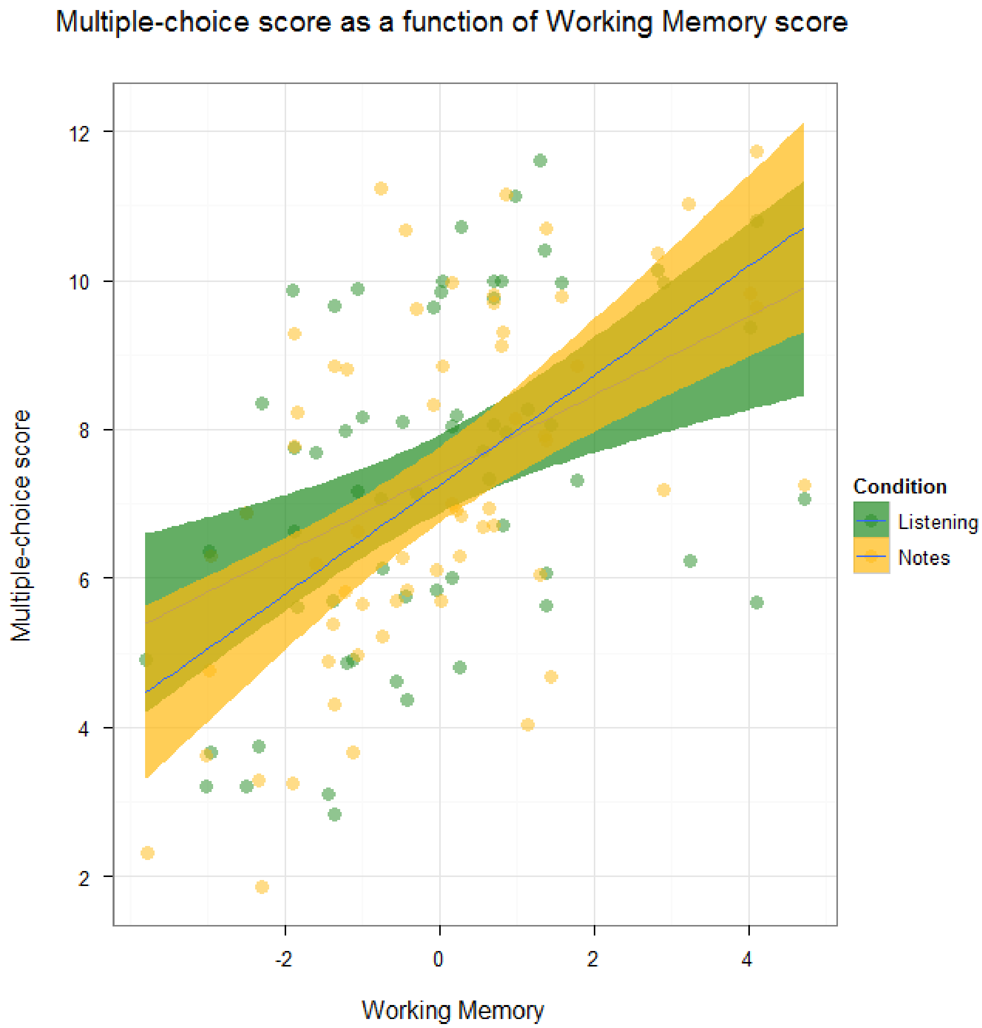

Figure 6. Multiple-choice scores as a function of working memory z-score in the *Listening Only* and *Allow Notes* conditions.

Impact of listening proficiency

Figure 7 shows performance on the multiple-choice assessments as a function of note taking condition (*Listening Only* vs. *Allow Notes*) and listening proficiency, as independently measured by the Versant™ listening subscore. Listening proficiency has the expected impact on performance such that the accuracy of a participant's performance on multiple choice increases as his/her listening proficiency increases (as measured by the Versant™). The figure also shows no clear interaction between listening proficiency and performance in the *Listening Only* and *Allow Notes* conditions.

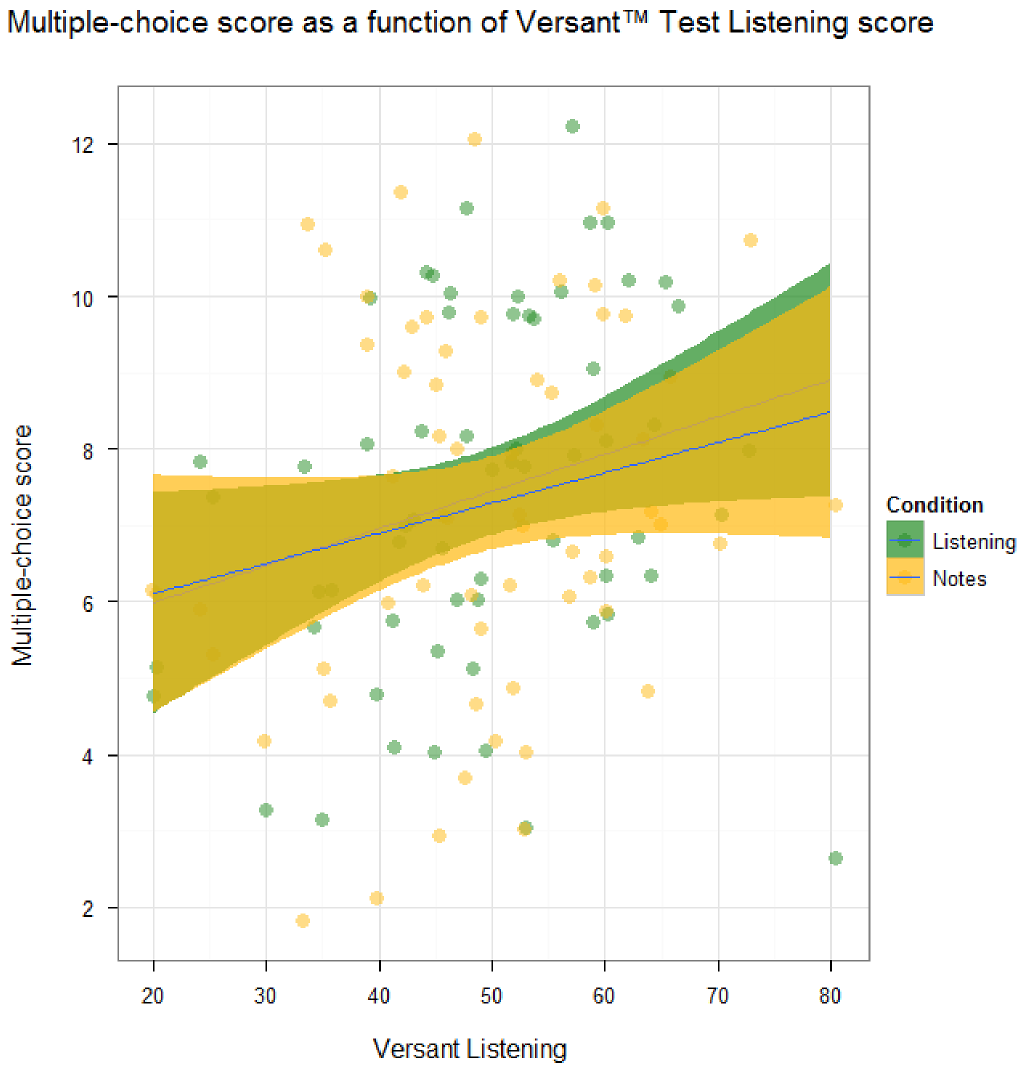

Figure 7. Multiple-choice scores as a function of Versant™ listening subscore in the *Listening Only* and *Notes Allowed* conditions.

These results indicate that while giving the listener the option to take notes does not change his/her performance when taking a DLPT-like multiple-choice listening proficiency test, the test does show differences when other measures, like working memory capacity and independently-measured listening proficiency, also *do* show an impact. That is, the fact that note taking had no effect is not because our test was insensitive to variables that have previously shown differences.

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). Items that received credit were either the target word, or a word that was an unambiguous misspelling of the target word (e.g., *againist*, *aganst*, *agiainst*, and *aginst* were accepted for the target word *against*), as did multiple-word responses that included either the target word or a misspelling of the target word (e.g., *Cosby's comments* received full credit for the target word *comments*). 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; 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*, *hapy*, 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, or blank responses.

Table 5 presents the average recall level for the participants out of 50 possible. The maximum number correct was 35 (out of 50), which indicates worse maximum performance than the maximum performance on multiple-choice questions (with a maximum of 24). These data did not suffer from a ceiling effect.

Table 5. The descriptive statistics for the number (percent) correct on recall questions (50 possible).

N	**Mean**	***SD***	**Minimum**	**Maximum**
62	16.4 (32.7%)	9.0 (17.9%)	0 (0.0%)	35 (70.0%)

Table 6 provides a summary of the mean percent correct across participants.

Table 6. Mean accuracy (number (percentage) correct on the recall questions) and SD as a function of note taking condition (*Listening Only, Allow Notes*).

Length	**Mean**	**SD**
Listening only	8.1 (32.3%)	4.9 (19.5%)
Allow notes	8.3 (33.2%)	4.6 (18.6%)

As can be seen in Figure 8, the mean score for the *Listening Only* condition was essentially the same as that of the *Allow Notes* condition. As with the multiple-choice data, these boxplots do not take into account the potential contributing effects of working memory and listener proficiency.

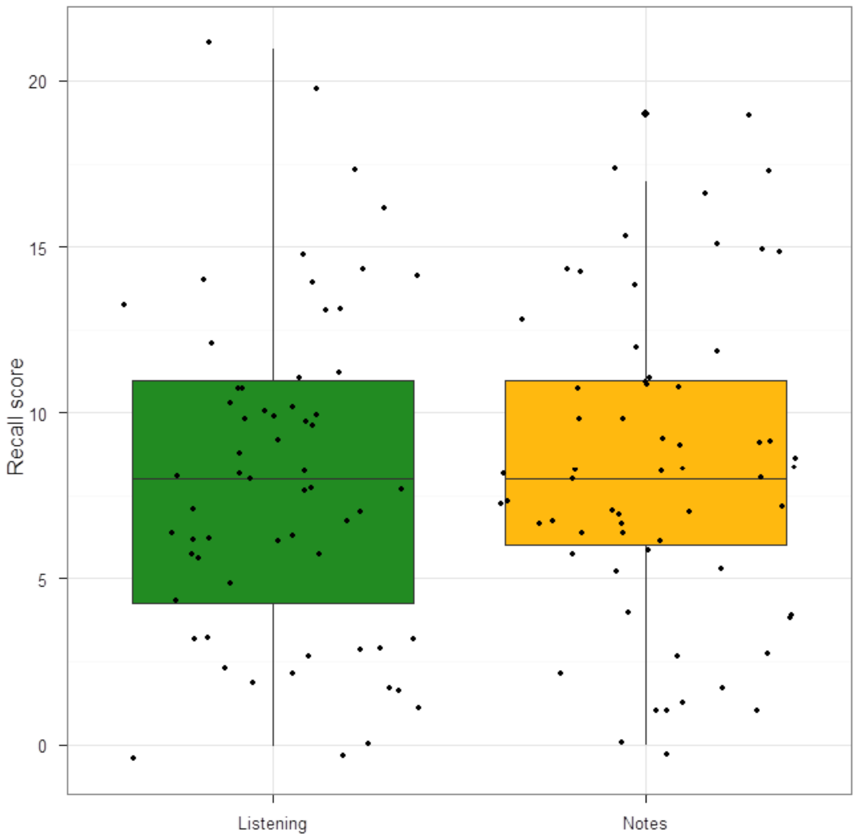

Figure 8. Recall scores by condition for all participants.

Our analysis used the same mixed-logit model as with the multiple-choice data. The results are shown in Table 7, below.

Table 7. Summary of the fixed effects in the mixed-logit model (Number of observations = 1488, log-likelihood = -856.7).

Fixed Effects	Estimate[1]	SE	z value	Pr(>|z|)	Odds
(Intercept)	-3.24	0.55	-5.89	0.00***	0.04
Note taking Condition (Allow Notes vs. Listening Only)	-0.01	0.10	-0.07	-0.07	0.04
Working Memory	0.43	0.07	6.38	0.00***	0.06
Versant™ Test Listening Score	0.04	0.01	4.12	0.01***	0.04

[1]The raw coefficients are expressed in logits (i.e., log-odds)
*** Significant at *p*-value less than 0.001

The Pr(>|z|) column indicates that there is still no evidence that the ability to take notes during a DLPT-like listening test offers an advantage to the test taker ($p > .05$). There continues to be evidence that performance changes (as expected) depending on the participant’s listening proficiency ($p = .01$) and working memory capacity ($p < .01$). If we had not seen these effects, it could have been the case that our measure was insensitive to the relevant differences. The null finding for note taking indicates that giving the listeners the option to take notes does not change their performance even when the task is more challenging, as in this recall test.

Impact of working memory

Figure 9 shows performance on the recall questions as a function of the normalized working memory z-scores and note taking condition (*Listening Only* vs. *Notes Allowed*). Working memory again has an enormous impact on performance such that the participant's performance accuracy on recall increases as working memory increases. In this case, there is absolutely no indication that taking notes helps people with more working memory capacity more than it helps those with less capacity.

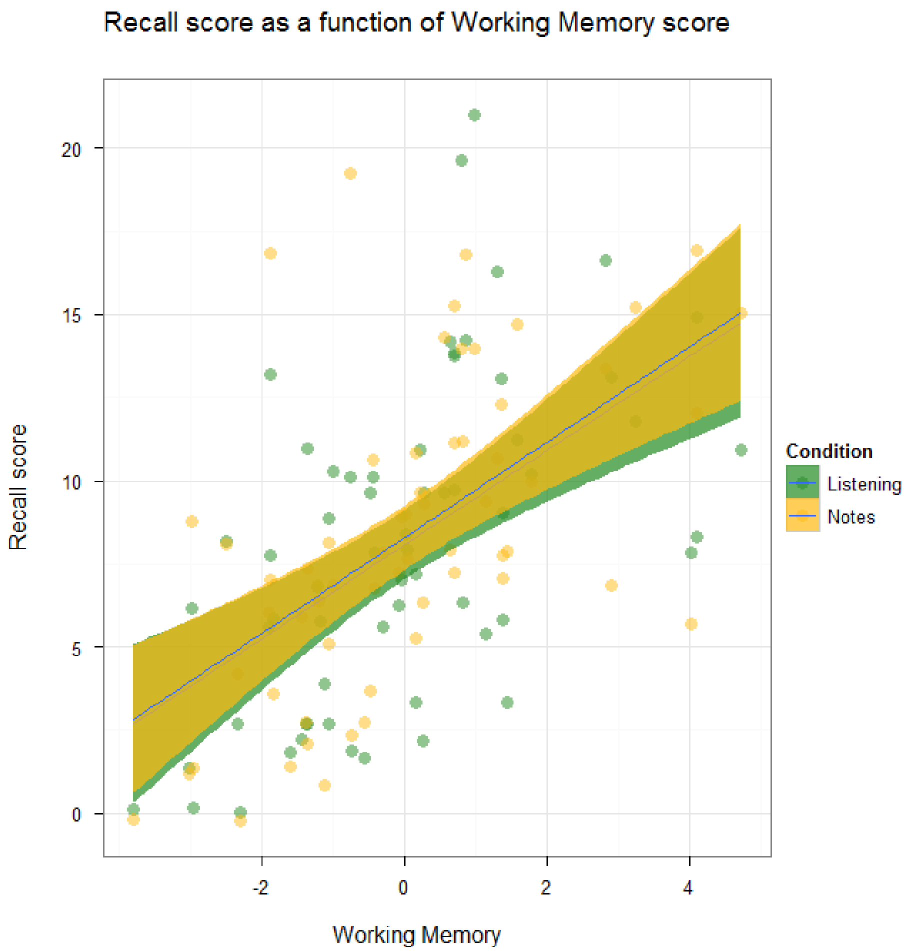

Figure 9. Recall scores as a function of working memory z-score in the *Listening Only* and *Allow Notes* conditions.

Impact of listening proficiency

Figure 10 shows performance on the recall task as a function of note taking condition (*Listening Only* vs. *Allow Notes*) and listening proficiency, as measured by the Versant™ listening subscore. Listening proficiency exhibited the expected impact on performance such that participant's performance accuracy on recall increases as listening proficiency increases. The figure also shows no clear interaction between listening proficiency and performance in the *Listening Only* and *Allow Notes* conditions.

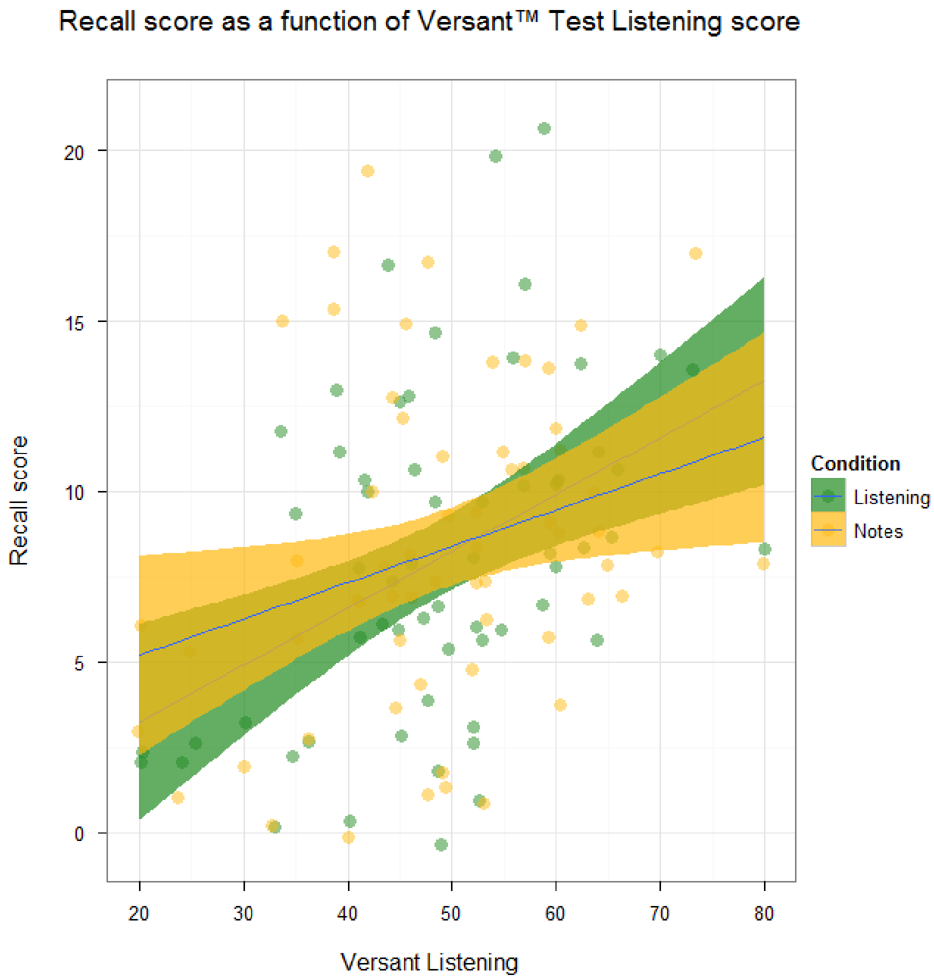

Figure 10. Recall scores as a function of Versant™ listening subscore in the *Listening Only* and *Notes Allowed* conditions.

These results indicate that while giving the listener the option to take notes does not change their performance when trying to remember words heard in a passage, other measures like working memory capacity and listening proficiency clearly do have an impact.

Further Analyses

Although the main analysis failed to find an effect for the note taking condition, not all of the participants actually took notes even when they had permission to do so in the *Allow Notes* condition. For this reason, we wanted to further explore the performance of participants who actually took notes when allowed.

Prolific Note Takers

For any given participant, there were six passages on which s/he would have been allowed to take notes. Table 8 shows the number of passages on which notes were taken by participants. Many participants (27 of the 62 tested) did not take notes on any passages even when allowed to do so, and only seven participants took notes on all six passages in the note taking condition. Table 8, below, shows the number of participants who took notes on some number of passages.

Table 8. Number of passages on which participants took notes in the note taking condition

# of Passages	0	1	2	3	4	5	6
# of Participants who took notes on that # of passages	27	13	4	3	1	7	7

It is unclear why 27 participants chose not to take any notes at all. On the post-test survey, almost half (48%) of the participants responded *Strongly agree* or *Agree* to the statement "Taking notes made answering the questions more difficult." However, 41% of participants also responded *Strongly agree* or *Agree* to the statement, "Taking notes helped me listen carefully to the passages," and 44% responded similarly to the statement "I felt more at ease when I could take notes than when I could not." Appendix G provides detailed results for all survey statements, and Appendix H shows the performance by condition for each response for all participants.

To further investigate the effect of note taking on listening comprehension, we decided to perform some exploratory descriptive analyses using the subset of participants who took the most notes – the "note takers." For our purposes, this subset consists of people who took notes on four or more of the passages, for a total of 15 "note takers." The mean Versant™ Listening score for the "note takers" was 48.2, which is very close to the overall mean of 49 across all participants.

Accuracy for "note takers" on the multiple-choice items is shown in Table 9, and in Figure 11; accuracy for the recall items is shown in Table 9 and in Figure 12. While accuracy is slightly higher in the note taking condition for note takers, the small sample size, rather large variation, and completely overlapping distributions indicate that this is not a significant trend.

Table 9. Mean number (percentage) correct and SD on multiple choice questions (12 total) and recall questions (25 total) as a function of notes condition for "note takers" (*Allow notes, Listening only).*

	Item Type			
	Multiple Choice		**Recall**	
Condition	**Mean**	**SD**	**Mean**	**SD**
Listening only	7.9 (66.1%)	1.9 (15.9%)	9.7 (38.7%)	5.1 (20.2%)
Allow notes	8.1 (67.8%)	2.2 (18.3%)	10.0 (40.0%)	5.1 (20.2%)

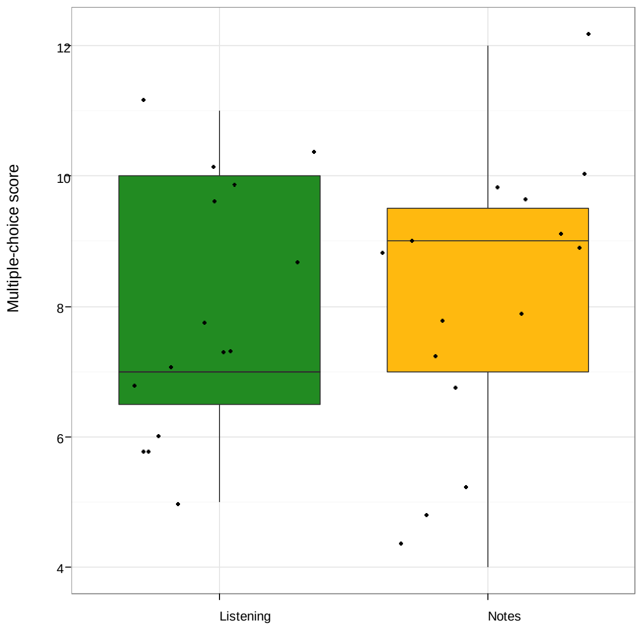

Figure 11. Multiple-choice scores by condition for note takers. Note takers are individuals who took notes on more than 4 (66%) of the passages in the Allow Notes condition.

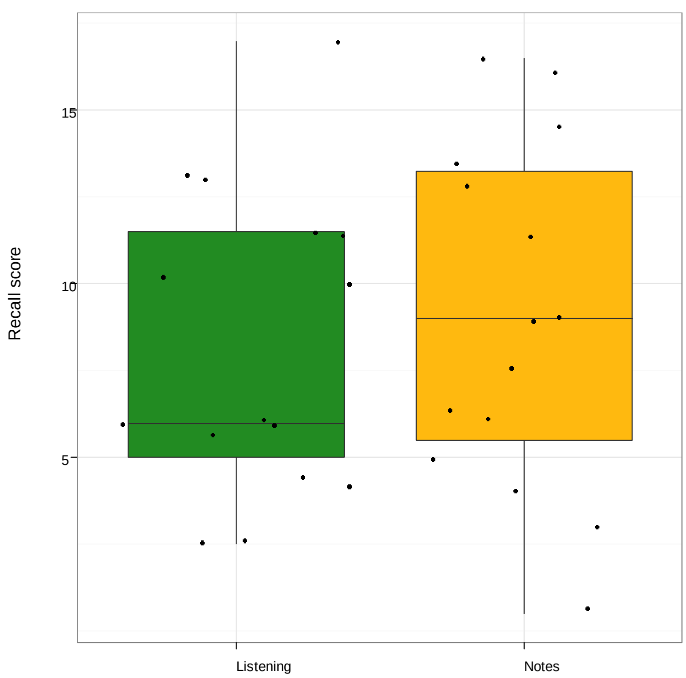

Figure 12. Recall scores by condition for note takers. Note takers are individuals who took notes on more than 4 (66%) of the passages in the Allow Notes condition.

Figures 13 and 14 corroborate this conclusion by showing the change in performance between conditions for all participants and then for the “note takers” only. These graphs show that while some participants perform better when they are allowed to take notes, other participants’ performance does not change across conditions, and others actually perform worse when they are allowed to take notes. This holds true even for our note takers (shown in red in the panel on the right side of Figures 13 and 14.)

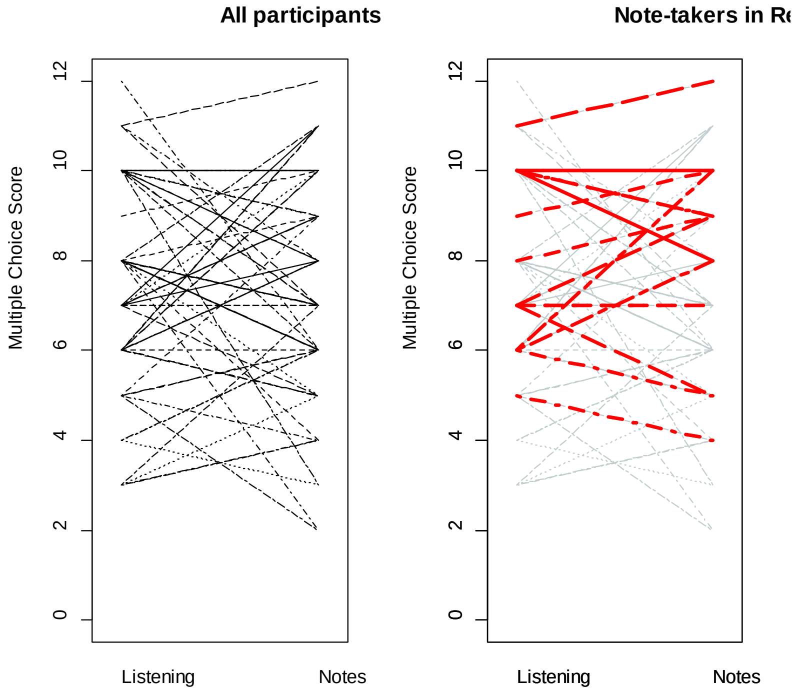

Figure 13. Multiple-choice score by condition. Note takers are individuals who took notes on more than 4 (66%) of the passages in the Allow Notes condition.

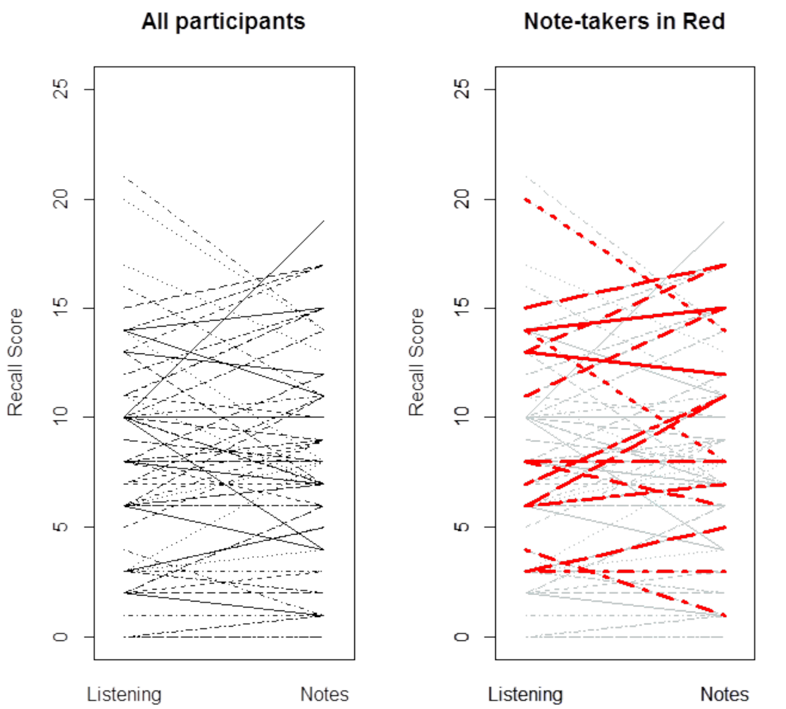

Figure 14. Recall score by condition. Note takers are individuals who took notes on more than 4 (66%) of the passages in the Allow Notes condition.

Quality of Notes

As part of our exploratory analyses, we coded notes for relevance. A relevant note was defined as containing information related to the main idea of the passage. We also coded notes for relevance to each of the multiple choice items. We found no significant correlation between the relevance of a note to the multiple choice question and the participant's performance on that item (r = .146, $p > .05$).

Keyboarding Ability

Because note taking in this study was done on the computer rather than on paper, participants who were better typists might have been more able to take advantage of the ability to take notes in the *Allow Notes* condition. Yet, in the questionnaire data, only just over a quarter (27%) of the participants answered "Strongly Agree" or "Agree" to the question "Typing notes prevented me from taking the notes I wanted."

Conclusions

Returning to our original questions, the results are clear.

- Allowing note taking has no impact on performance on a DLPT-like L2 listening comprehension test.
- Even participants who actually take notes when they are allowed to do not perform better.

While it may be the case that people who like to take notes while listening feel constrained by testing conditions that do not allow them to do so, there is no evidence that taking notes is advantageous when the multiple-choice test items are shown prior to playing an audio passage and remain on the screen while the passage is playing, as is the case on the DLPT. It is important to note that this test condition is very different than the typical work environment, where people may listen to a passage without knowing what it is that they are listening for. That environment is more similar to the recall task in which participants had no prior knowledge about which words would be important to remember. Yet even in that case, the ability to take notes did not alter performance. For this reason, test developers who are designing DLPT-like tests can feel confident that prohibiting test-takers from taking notes during the test will have no impact on test scores.

It is not the case that our test was insensitive to differences. Participant's listening proficiency in the language, as measured independently by the Versant™, clearly predicted performance, as did working memory. It was our original hope that by factoring out these well-known contributions to performance differences, we would enhance the likelihood of finding effects of note taking. But even with these variables accounted for, our participants performed no differently when they were and were not allowed to take notes.

SUMMARY

CASL researchers investigated the effect of note taking on a DLPT-like listening test by conducting a research study in which participants answered test questions in both an *Allow Notes* and a *Listening Only* condition. Empirical results were analyzed with a linear mixed effects approach in which the note taking condition was treated as a fixed effect, and items and participants were treated as random effects. Working memory and listening proficiency scores were included as covariates. Permitting participants to take notes did not result in better performance than forbidding them to take notes; there is no evidence that allowing notes on a DLPT-like listening test alters performance in a systematic way. Thus, we have found no evidence that would suggest that allowing note taking would impact performance on the DLPT.

REFERENCES

Atkins, S. M., Harbison, J. I., Bunting, M. F., & Dougherty, M. R. (2012). Using online games to assess cognitive ability: Validity of two web-deployable measures of cognitive ability.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829–839.

Barbier, M.-L., & Piolat, A. (2005). L1 and L2 cognitive effort of note taking and writing. InL. Allal, Dolz, J., & Rochat, F. (Eds.). *Proceedings writing 2004 9th International Conference of the EARLI Special Interest Group on Writing, University of Geneva, September 20-22, 2004*. Geneva, Switzerland: University of Geneva.

Bates, D.M., Maechler, M., & Bolker, B. (2012). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

Bernstein, J., & De Jong, J. H.A.L. (2001). An experiment in predicting proficiency within the Common Europe Framework Level Descriptors. In Y.N. Leung et al. (Eds.), *Selected Papers from the Tenth International Symposium on English Teaching* (pp. 8-14). Taipei, ROC: The Crane Publishing.

Blodgett, A., Bloomfield, A., Wayland, S., O'Connell, S. Linck, J., Gynther, K., & Kramasz, D. (2011). Second language listening comprehension: Assessment plan (Report No. 81434-C-1.5.5.2.). College Park, MD: University of Maryland, Center for Advanced Study of Language.

Bloomfield, A., Wayland, S. C., Rhoades, E., Blodgett, A., Linck, J., & Ross, S. (2011). *What makes listening difficult?* (TTO 81434 Technical Report E.3.1). College Park, MD: University of Maryland Center for Advanced Study of Language.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International 5(9/10), 341–345.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: SAGE.

Carrell, P. L. (2007). *Notetaking strategies and their relationship to performance on listening comprehension and communicative assessment tasks*. (TOEFL Monograph Series No. MS-35). Princeton, NJ: Educational Testing Service.

Carrell, P.L., Dunkel, P. A., & Molluan, P. (2002). *The effects of notetaking, lecture length and topic on the listening component of the TOEFL 2000*. (TOEFL Monograph Series No. MS-23). Princeton, NJ: Educational Testing Service.

Chaudron, C., Cook, J., & Loschky, L. (1988). Quality of lecture notes and second language listening comprehension (Tech. Rep. No. 7). Honolulu: University of Hawaii at Manoa, Center for Second Language Classroom Research.

Chaudron, C., Loschky, L., & Cook, J. (1994). Second language listening comprehension and note-taking. In J. Flowerdew (Ed.) Academic Listening: Research Perspectives (pp. 75–92). Cambridge: Cambridge University Press.

Clark, M., Wayland, S., Castle, S., & Gynther, K. (2013). The effects of note-taking on L2 listening comprehension: Assessment plan (TTO 2012 Technical Report 2.1). College Park, MD: University of Maryland Center for Advanced Study of Language.

Clerehan, R. (1995). Taking it down: Notetaking practices of L1 and L2 students. *English for Specific Purposes*, *14*(2), 137-155.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100.

Crawford, C. C. (1925a). The correlation between college lecture notes and quiz papers. *The Journal of Educational Research*, *12*(4), 282-291.

Crawford, C. C. (1925b). Some experimental studies of the results of college note-taking. *The Journal of Educational Research*, *12*(5), 379-386.

Cunnings, I. (2012). An overview of mixed-effects statistical models for second language researchers. *Second Language Research*, *28*, 369--382.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning & Verbal Behavior*, *19*, 450–466.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1—28.

Derwing, T., & Munro, M. (2001). What speaking rates do non-native listeners prefer? Applied Linguistics, 22(3), 324–337.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: with the lme4 package. *Journal of Statistical Software*, *20*(2), 1—18.

Dunkel, P. (1988). The content of L1 and L2 students' lecture notes and its relation to test performance. *TESOL Quarterly*, *2*(2), 259-281.

Dunkel, P., Mishra, S., & Berliner, D. (1989). Effects of note taking, memory, and language proficiency on lecture learning for native and nonnative speakers of English. *TESOL Quarterly*, *23*(3), 543–549.

Gold, J. M., Carpenter, C., Randolph, C., Goldberg, T. E., & Weinberger, D. R. (1997). Auditory working memory and Wisconsin Card Sorting Test performance in Schizophrenia. Archives of General Psychiatry, 54, 159–165.

Hale, G. A., & Courtney, R. (1994). The effects of note-taking on listening comprehension in the Test of English as a Foreign Language. *Language Testing*, *11*(1), 29–47.

Hartley, J. (1983). Notetaking research: Resetting the scoreboard. *Bulletin of the British Psychological Society*, *36*, 13–14.

Hartley, J. (2002) Notetaking in non-academic settings: A review. *Applied Cognitive Psychology*, *16*, 559-574.

Harrington, M., & Sawyer, M. (1992). L2 working memory capacity and L2 reading skill. Studies in Second Language Acquisition,14(1), 25–38.

Hayati, A. M., & Jalilifar, A. (2009). The impact of note-taking strategies on listening comprehension of EFL learners. *English Language Teaching*, *2*(1), 101-111.

Jaeger, T. F. (2008). Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434-446.

Jacewicz, E., Fox, R. A., O'Neill, C., and Salmons, J. (2009). Articulation rate across dialect, age, and gender. *Language Variation and Change,* 21(2): 233-256.

Kiewra, K. (1985). Investigating notetaking and review: A depth of processing alternative. *Educational Psychologist*, *20*(1), 23-32.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573—603.

Levin, I. P. (1999). Relating statistics and experimental design: An introduction. Thousand Oaks, CA: SAGE.

Lin, M. (2006). The effects of note-taking, memory and rate of presentation on EFL learners' listening comprehension. Unpublished doctoral dissertation, La Sierra University, California.

Marian, V., Blumenfeld, K., & Kaushanskaya, M. (2007). Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multi-linguals. *Journal of Speech Language and Hearing Research, 50* (4), 940–967.

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. Journal of Memory and Language, 55(3), 381–401.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy, & L. E. Bourne (Eds.), Foreign language learning (pp. 339–364). London: Lawrence Erlbaum Associates.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. Speech Communication, 9(5-6), 453–467.

Pearson (2010). *Versant™ Pro—Speaking*. Technical Paper. 1–10. Pearson Education, Inc. Retrieved from http://www.versanttest.com/products/proSpeaking.jsp

Piolat, A., Olive, T., & Kellogg, R. T. (2005). Cognitive effort during note taking. *Applied Cognitive Psychology*, *19*(3), 291–312.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornick, F. Leisch, & A. Zeileis (Eds.) *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, March 20-22, Vienna, Austria.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Rosenhouse, J., Haik, L., & Kishon-Rabin, L. (2006). Speech perception in adverse listening conditions in Arabic-Hebrew bilinguals. *International Journal of Bilingualism*, *10*(2), 119–135.

Sprenger, A.M., Atkins, S.M., Colflesh, G.J.H., Briner, T.L., Buchanan, J.B., Chavis, S.E., Chen, S.Y., Iannuzzi, G.L., Kashtelyan, V., Dowling, E., Bolger, D.J., Bunting, M.F., & Dougherty, M.R. (in preparation). A four-dimensional video game for measuring cognitive ability.

Song, M. (2012). Note-taking quality and performance on an L2 academic listening test. *Language Testing*, *29*(1), 67-89.

Tsai, T., & Wu, Y. (2010). Effects of note-taking instruction and note-taking languages on college EFL students' listening comprehension. *New Horizons in Education*, *58*(1), 120-132.

Wayland, S., O'Connell, S., Linck, J., Kramasz, D., Gynther, K., Bloomfield, A., Blodgett, A., Silbert, N., & Saner, L. (2012). Second language listening comprehension: The impact of speech rate, passage length, and information density (TTO 2001 Technical Report 1.1). College Park, MD: University of Maryland Center for Advanced Study of Language.

APPENDIX A: PASSAGE MANIPULATION

Although the DLPT test specifications provide general guidance on selecting passages appropriate for testing various ILR levels, passages derived from authentic sources can still vary considerably in terms of speech rate, density of information, and other factors. In order to minimize the effect of passage-specific variation, we used passage versions that had been modified to narrow the range of speech rate, information density, and passage length. Without controlling for passage length, density, and speech rate, it would be difficult to disentangle the variation in performance due to note taking from the variation in performance due to differences in passage characteristics. These manipulations allowed us to more clearly see an interaction between note taking and language proficiency level. The manipulated passages used in this study represent a subset of those developed for CASL's empirical investigation into factors affecting listening proficiency (Blodgett, Bloomfield, Wayland, O'Connell, Linck, Gynther, & Kramasz, 2011). None of these manipulations change the basic mode of the passage (e.g., instructive, evaluative).

Length and Density

We intentionally manipulated and controlled passage length at each level to try to minimize its effect. The original passages varied widely in length, so length was standardized within each ILR level by adding non-redundant content when the source passage was too short or by removing content when the source passage was too long. All length manipulations were reviewed by team members in order to maintain coherence. As seen in Table 1, for level 2, passage length, measured in syllables, ranges from 218 to 236 with an average of 226 syllables. This corresponds to audio cut lengths between 0:54 minutes and 1:15 minutes, which is generally in line with the DLPT specifications for level 2. The range for level 3 passages was between 355 syllables and 363 syllables, with an average of 358 syllables. This corresponds to recorded lengths between 1:34 minutes and 1:40, which is consistent with DLPT specifications for level 3.

Lexical density is typically measured by the type/token ratio.[8] This is the ratio of the number of unique words (*types*) in the passage to the total number of words (*tokens*). A passage with a higher type/token ratio can be considered to have more lexical variety than a passage with a lower ratio. Because words can occur in many forms (e.g., *dog/dogs, help/helping*), we employed the following rules, primarily drawing on Richards (1987), in determining which lexical items would be classified as unique types:

- Hyphenated words count as one word
- Inflected and non-inflected forms of the same stem count as a single type (e.g., *go*, *goes*)
- Contractions of subject and predicate (e.g., *I'm*, *he's*) are treated as two words. Full and contracted forms 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.
- Interjections (e.g., *oh*), hesitations (e.g., *ah*, *um*), and false starts are included.
- No distinction between verb forms in their main and auxiliary role (e.g., *do*)

For this study, we implemented a measure of type/token ratio developed by Covington & McFall (2010) called the *Moving Average Type-Token Ratio*, or MATTR. Rather than using the passage as a whole, which tends to decrease the type/token ratio for longer passages, this method splits the passage into successive *windows* of a fixed token length and calculated the type/token ratio independently for each window. Because this measure averages over successive window sizes smaller than the length of the passage, it is less sensitive to passage length. Type/token ratios of the passages were standardized to have a value between .69 and .73. In most cases, this meant increasing the type/token ratio of the original passage. To do this, we replaced words with tokens of previously unused types (e.g., *helping* was replaced with *assisting* if the verb *help* had previously been used in the passage).

[8] We recognize that measuring density in this way does not capture the impact of a speaker's intentional use of specific lexical items and that passages of similar lexical density can vary greatly in this regard.

Speech Rate

Empirical studies suggest that normal speech rate is roughly five syllables per second (e.g., Jacewicz, Fox, O'Neill, & Salmons, 2009,; Derwing & Munro, 2001). However, the speech rates for our original source passages was lower, and varied from 3.4 syllables/second to 5.2 syllables/second, with an average of 3.9 syllables/second. Level 3 source passages did not necessarily have higher speech rates than level 2 passages. To standardize the speech rate across passages, the articulation rate was manipulated using the Pitch Synchronous Overlap and Add (PSOLA) algorithm (Moulines & Charpentier, 1990) in Praat software (Boersma, 2001) to be 4.0 syllables/second for all passages.[9] This manipulation changes the rate of speech without altering the perceived pitch.

Passage Recordings

Transcriptions of the original passages were used as a base when creating the manipulated (long length, high density) versions using the process outlined above. These modified transcriptions were then recorded by professional voice actors, using the original recordings as a reference for their performances. To the extent possible, the non-linguistic components of the original recordings (e.g., background noise, introductory theme music) were recreated in these rerecorded versions. The length and density manipulations were part of the recording script; the speech rate manipulations were performed directly on the recordings. The resulting recordings were reviewed by CASL researchers to ensure that the manipulations did not sound artificial and that the voice performances matched the original recordings to the greatest extent possible.

[9] The articulation rate is the rate of the actual stretches of speech, not including pauses between speakers. A rate of 5.0 syllables/second was initially tried, but seemed unnaturally fast to native speakers.

APPENDIX B: SAMPLE TEST PASSAGES AND QUESTIONS

Sample Passage Transcript:

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.

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.

Sample Recall Items:

Keyword: *pushed*
Dr. Bekoff told us how the coyote can survive the kind of persecution that _______ the bison, wolf, and passenger pigeon to extinction.

APPENDIX C: MODIFIED VERSION OF THE LEAP-Q

Based on LEAP-Q from Marian, Blumenfeld, and 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: _________________________
__learning disability: __________________________
[Page 2] (*participants complete the same set of questions for each language they listed above*).

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

Speaking	0	1	2	3	4	5	6	7	8	9	10
Understanding spoken language	0	1	2	3	4	5	6	7	8	9	10
Reading	0	1	2	3	4	5	6	7	8	9	10

(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

Interacting with friends	0	1	2	3	4	5	6	7	8	9	10
Language tapes/self instruction	0	1	2	3	4	5	6	7	8	9	10
Interacting with family	0	1	2	3	4	5	6	7	8	9	10
Watching TV	0	1	2	3	4	5	6	7	8	9	10
Reading	0	1	2	3	4	5	6	7	8	9	10
Listening to the radio	0	1	2	3	4	5	6	7	8	9	10

(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

Interacting with friends	0	1	2	3	4	5	6	7	8	9	10
Language tapes/self instruction	0	1	2	3	4	5	6	7	8	9	10
Interacting with family	0	1	2	3	4	5	6	7	8	9	10
Watching TV	0	1	2	3	4	5	6	7	8	9	10
Reading	0	1	2	3	4	5	6	7	8	9	10
Listening to the radio	0	1	2	3	4	5	6	7	8	9	10

APPENDIX D: NOTE TAKING QUESTIONNAIRE

Please answer the following questions about your experience with note taking in general. Read each of the following statements and indicate whether you agree or disagree with the statement. Circle the number (5, 4, 3, 2, or 1) that best describes your opinion about the statement.

5 = Strongly Agree **4** = Agree **3** = Neither agree nor disagree **2** = Disagree **1** = Strongly Disagree

	Strongly Agree		Neither agree nor disagree		Strongly Disagree
1. I have had training in developing note taking skills.	5	4	3	2	1
2. I often take notes when I am listening to information in English.	5	4	3	2	1
3. I often take notes when I am listening to information in my native language (Spanish).	5	4	3	2	1

Please answer the following questions about your experience taking this test. Read each of the following statements and indicate whether you agree or disagree with the statement. Circle the number (5, 4, 3, 2, or 1) that best describes your opinion about the statement.

5 = Strongly Agree **4** = Agree **3** = Neither agree nor disagree **2** = Disagree **1** = Strongly Disagree

	Strongly Agree		Neither agree nor disagree		Strongly Disagree
1. Taking notes helped me to answer the questions better than if I had not been able to take notes.	5	4	3	2	1
2. I felt more at ease when I could take notes than when I could not.	5	4	3	2	1
3. Taking notes made answering the questions more difficult.	5	4	3	2	1
4. Taking notes helped me listen carefully to the passages.	5	4	3	2	1
5. Taking notes helped me to understand the passages.	5	4	3	2	1
6. Taking notes distracted me from paying attention to the information in the passages.	5	4	3	2	1
7. I had enough time to take as many notes as I wanted.	5	4	3	2	1
8. The passages were too short for note taking to help me very much.	5	4	3	2	1
9. The passages didn't contain enough information for note taking to help me very much.	5	4	3	2	1
10. The speakers in the passage were speaking too slowly for note taking to help.	5	4	3	2	1
11. Typing the notes prevented me from taking the notes I wanted.	5	4	3	2	1
12. The passages were too hard for note taking to help	5	4	3	2	1

me very much.					
13. The speakers in the passage were speaking too fast for note taking to help very much.	5	4	3	2	1
14. Note taking would have helped more if I didn't know the test questions ahead of time.	5	4	3	2	1

Partially adapted from Carrel et al. (2002)

APPENDIX E: DEMOGRAPHIC DESCRIPTORS

Table E.1. Participant Demographics

	Age (N=60, 2 omitted*)	Length of Residence in L2 Country (in years) (N=46, 16 omitted*)	Time spent in a school/work environment where English is used (in years) *(N=45, 17 omitted*)*
Mean *(SD)*	37.2 (11.94)	12.5 (9.5)	7.34 (6.9)
Median	38	10.54	6
S.E.	1.54	1.24	1.03
Min	19	0.33**	0
Max	65	40.25	29

*Omitted participants did not respond to questionnaire item
**Decimals represent months reported by participants

Table E.2. Highest Level of Education Attained

	Frequency	Percent	Cumulative Percent
Less than High School	2	0.03	0.03
High School	15	0.24	0.27
Professional Training	2	0.03	0.3
Some college	12	0.19	0.49
College	13	0.21	0.7
Some Graduate School	4	0.06	0.76
Masters	5	0.08	0.84
Ph.D./M.D./J.D.	0	0	0.84
Other	3	0.05	0.89
(missing)	6	0.1	0.99
Total	62	1	

APPENDIX F: BAYESIAN ANALYSIS OF MULTIPLE-CHOICE RESULTS

In empirical assessments, we are usually not interested in the specific statistics of our measured participants, but rather in how those measured data relate to the underlying population parameters of our construct of interest. Bayesian estimation is a method that allows for the "reallocation of credibility across a space of candidate possibilities" (Kruschke, 2013, p. 574) toward parameter values that are consistent with the data. Rather than providing a single test of significance based on comparison with sampling distributions from a null hypothesis, Bayesian estimation generates the distribution of credible parameter estimates given the data. The interpretation of the meaning of these credible parameter estimates is left to the researcher and is not dependent solely on reaching a critical *p*-value. Bayesian analysis thus allows you to test the null hypothesis. If you find that most of your data falls within the Region of Practical Equivalence (ROPE; see below for an explanation), you can safely conclude that there truly is no difference between your conditions. For a study like this one, where the results indicate that there are no differences between conditions, testing the null hypothesis helps us to interpret our data.

For this reason, as a complement to our main analysis, we decided to perform an exploratory assessment of our note taking data from a Bayesian perspective. This exploration is essentially a Bayesian version of what would be a repeated-measures *t*-test in the null hypothesis significance testing (NHST) paradigm. For each participant, a difference score was calculated between the scores for the two conditions (Listening Only – Allow Notes). If the scores for the two conditions were similar, one would expect these difference scores to be close to 0. A greater score in the *Allow Notes* condition would lead to a negative difference score. The Bayesian approach allows us to see the credible parameter estimates for these difference scores.

The analysis was run in R statistical software (R Core Team, 2013) using the JAGS sampler (Plummer, 2003) for the Monte-Carlo Markov Chain (MCMC) simulation. The code used for the analysis[10] was developed by John Kruschke based on the approach detailed in Kruschke (2013). The program uses 100,000 steps in MCMC sample, with 500 steps used to tune the sampler and 1000 for burn-in. Initial values for the MCMC chains are set based on the data (i.e., *mu* is set using the data mean and *sigma* is set using the data standard deviation). Running the R code generates a number of useful graphical representations of the results, which are presented in Figure E.1.

The box on the upper right of the figure shows a histogram of the data (in red) overlaid with a small sampling of the thousands of generated parameter values that are compatible with the data. All of the other histograms in the figure show the posterior distributions -- the probability density of those generated parameter values.

The histogram in the top left corner shows the distribution of the means of the generated distributions. That is, after generating tens of thousands of distributions consistent with the data, the mean of that set of parameter values is 0.139. A thick black line at the base of the histogram shows the 95% Highest Density Interval (HDI); points within this range have a higher probability than points outside of it. Thus, the "most credible" parameter values, given our data, fall within this range. To aid in interpretation, we have illustrated a Region of Practical Equivalence (ROPE) in red that extends from -0.5 to 0.5. That is, we are willing to accept values in the ROPE as being equivalent for practical purposes. Setting the ROPE to these values is a conceptual, not statistical, choice; we believe that an average of half a point difference or less between the two conditions is practically equivalent to no difference.[11] Eighty-one percent of the 95% HDI values fall within this ROPE. Had 100% of the values fallen within the ROPE, we would be in a position to essentially "accept" the null hypothesis of no difference between the two conditions.[12]

---

[10] R code available from: http://www.indiana.edu/~kruschke/DoingBayesianDataAnalysis/Programs/BEST1G.R

[11] The reader, of course, is free to disagree with this choice.

[12] Note that had we chosen a range of -1.0 to 1.0 for the ROPE, we could "accept" the null, as all of the 95% HDI would have fallen within that range. Alternatively, had we chosen a value of -.01 to .01 for the ROPE, a much smaller percent of the values would have fallen within that range. Rather than being a problem for this type of analysis, it serves as a reminder that statistical information cannot be divorced from practical considerations when assessing its implications.

The other histograms illustrate similar information for the standard deviation, effect size, and normality. This exploratory Bayesian analysis confirms that there is no meaningful effect for the Allow Notes condition in this study, and the credible parameter values for the mean difference between the conditions hover around 0.

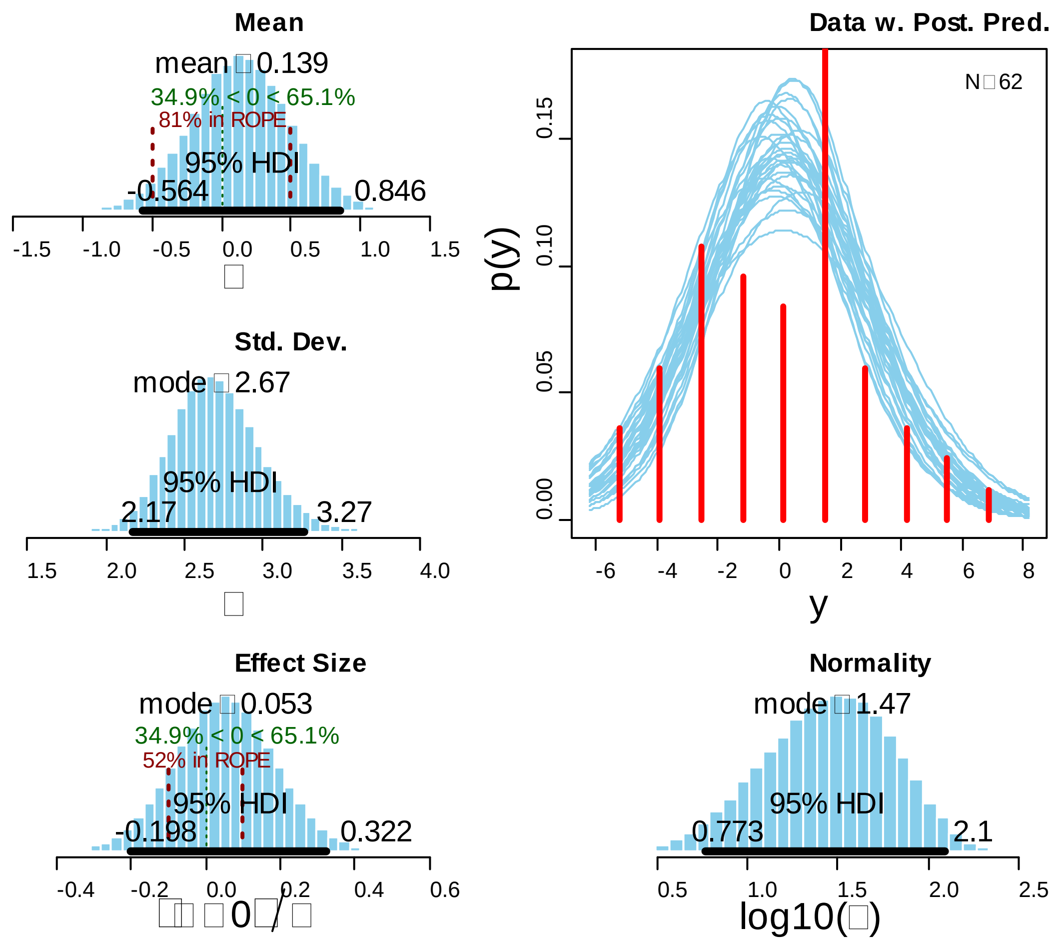

Figure E.1. Graphical output generated from Bayesian analysis.

APPENDIX G: SUMMARY OF NOTE TAKING QUESTIONNAIRE RESULTS

		Response						
Item	Statement	1	2	3	4	5	N	Mean
GQ1	I have had training in developing note taking skills.	8 (13%)	15 (24%)	18 (29%)	11 (18%)	10 (16%)	62	3.00
GQ2	I often take notes when I am listening to information in English.	8 (13%)	11 (18%)	18 (30%)	14 (23%)	10 (16%)	61	3.12
GQ3	I often take notes when I am listening to information in my native language (Spanish).	18 (29%)	11 (18%)	12 (19%)	13 (21%)	8 (13%)	62	2.74
SQ1	Taking notes helped me to answer the questions better than if I had not been able to take notes.	13 (21%)	8 (13%)	15 (25%)	14 (23%)	11 (18%)	61	3.03
SQ2	I felt more at east when I could take notes than when I could not.	19 (31%)	8 (13%)	11 (18%)	11 (18%)	12 (20%)	61	2.82
SQ3	Taking notes made answering the questions more difficult.	12 (20%)	13 (21%)	19 (31%)	10 (16%)	7 (11%)	61	2.79
SQ4	Taking notes helped me listen carefully to the passages.	9 (15%)	12 (20%)	17 (28%)	9 (15%)	14 (23%)	61	3.12
SQ5	Taking notes helped me to understand the passages.	12 (20%)	13 (21%)	14 (23%)	12 (20%)	10 (16%)	61	2.92
SQ6	Taking notes distracted me from paying attention to the information in the passages.	20 (33%)	9 (15%)	15 (25%)	12 (20%)	5 (8%)	61	2.56
SQ7	I had enough time to take as many notes as I wanted.	8 (13%)	9 (15%)	15 (25%)	13 (21%)	16 (26%)	61	3.24
SQ8	The passages were too short for note taking to help me very much.	8 (13%)	8 (13%)	20 (33%)	14 (23%)	10 (17%)	60	3.17
SQ9	The passages didn't contain enough information for note taking to help.	3 (5%)	13 (21%)	23 (38%)	12 (20%)	10 (16%)	61	3.21
SQ10	The speakers in the passages were speaking too slowly for note taking to help much.	8 (13%)	6 (10%)	16 (27%)	20 (33%)	10 (17%)	60	3.30
SQ11	Typing the notes prevented me from taking the notes I wanted.	13 (22%)	3 (5%)	17 (28%)	17 (28%)	10 (17%)	60	3.13
SQ12	The passages were too hard for note taking to help me very much.	9 (15%)	14 (23%)	20 (33%)	11 (18%)	6 (10%)	60	2.85
SQ13	The speakers in the passages were speaking too fast for note taking to help very much.	14 (23%)	10 (16%)	20 (33%)	6 (10%)	11 (18%)	61	2.84
SQ14	Note taking would have helped more if I didn't know the test questions ahead of time.	16 (26%)	6 (10%)	21 (34%)	6 (10%)	12 (20%)	61	2.87

Response key: 1 = Strongly Agree, 2 = Agree, 3 = Neither agree nor disagree, 4 = Disagree, 5 = Strongly Disagree; Questions GQ1 – GQ3 are about note taking in general, questions SQ1 – SQ14 are about note taking during this study. All questions were presented to the participants in Spanish

Corresponding Author and Reprints:
Martyn Clark, PhD, University of Maryland Center for Advanced Study of Language, (301) 226-8864, mclark@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 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 Jeffrey Crowson, PhD, Professor – Education Research; Defense Language Institute Foreign Language Center, (831) 242-3788, Jeffrey.j.crowson.civ@mail.mil.

