LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Using Utterance Recall to Assess Second Language Proficiency

Gabriel Culbertson, Erik Andersen,
and Morten H. Christiansen

Cornell University

Abstract: Obtaining quick and reliable evidence regarding the proficiency of learners is a perennial issue in second language (L2) learning research. In this study, we examined naturalistic utterance recall as a measure of L2 learning proficiency that can be easily extracted from videos and automatically scored using the video's captions. In our recall task, learners listen to audio clips and write down as much of the utterance as they can remember. We evaluated this naturalistic recall task with a sample of English native speakers who are learning Spanish at beginner to advanced levels, as well as Spanish native speakers. The results suggest that our recall measure is a better predictor of a learner's ability to translate heard sentences than a shortened version of a standardized listening multiple-choice comprehension test. Our findings suggest naturalistic utterance recall can offer an accurate and efficient method for predicting foreign language proficiency.

Keywords second language learning; second language proficiency; recall; memory; elicited imitation; sentence repetition

Introduction

Educators often need to assess second language (L2) learners' proficiency, for example, to place students at the right level in language classes, or to determine whether students may be able to follow instruction in a foreign language when

We'd like to thank our research assistants Justin Tung, Dana Daniels, Kiana Vidal, and Liane Longpre for their help in collecting materials for the study and analyzing the data. We'd also like to thank Solace Shen for providing feedback on the study design and survey as well as Erin Isbilen, Priscila López-Beltrán Forcada, and Manuel Pulido-Azpíroz for comments on a previous version of this paper.

Correspondence concerning this article should be addressed to Morten H. Christiansen, Department of Psychology, 228 Uris Hall, Cornell University, Ithaca, NY 14853, USA. E-mail: christiansen@cornell.edu

DOI: 10.1111/lang.12399

studying abroad. Accurate assessments of proficiency are also important for researchers seeking to understand the complex nature of L2 learning as well as to those seeking to evaluate new language learning paradigms and methodologies. Unfortunately, self-reported proficiency tends to be unreliable, and L2 proficiency tests are difficult to develop and require an extensive time commitment by learners to complete. For example, the national Japanese language proficiency test (JLPT) requires 3 hours to complete[1] and the National Spanish Examination (NSE) requires around an hour and a half.[2] Furthermore, it is often not possible to use these tests (e.g., JLPT and NSE) in research or classroom contexts because the exams are only permitted to be administered by specific testing organizations (though practice samples are sometimes available).

The challenge of assessing general L2 proficiency is further exacerbated because the various activities a learner may engage in might require quite different sets of vocabulary and even grammatical structures. Much of the vocabulary a learner will need for any specific activity may be limited to a small subset of situations. For example, in the commonly used Japanese textbook, *Genki*, learners acquire about 1,500 words,[3] whereas watching a season of a single television show like *Tokyo Diner* will require about 3,000 words.[4] Of the words learned in *Genki*, only about 600 are used in the *Tokyo Diner* television program. Thus measuring proficiency using simple knowledge tests like vocabulary translation tests may be measuring context-specific knowledge rather than overall language proficiency.

In two key reviews, Thomas (1994, 2006) noted a lack of consistency in measurements of proficiency in L2 learning research. Two more recent reviews by Tremblay (2011) and Hulstijn (2012) further highlighted the need for improvements in assessments of L2 proficiency (see Bowden, 2016, for updates). Currently, researchers often must resort to other less-than-ideal L2 proficiency measures. For example, many studies use vocabulary or pronunciation tests. However, language proficiency is not determined by recognizing or pronouncing isolated words. In real-life language situations, we encounter continuous streams of speech, often comprising multiple sentences, coming at us at a rate of about 300–350 syllables per minute (Studdert-Kennedy, 1986). That is, language happens in the here-and-now: we must rapidly process sounds, words, and other units or else incoming information is quickly lost (Christiansen & Chater, 2016). From this perspective, Christiansen and Chater argue that rapid chunking of language input is central to language proficiency. Though there has been some work using chunking to measure general skill learning (e.g., Isbilen, McCauley, Kidd, & Christiansen, 2017, in press—see Christiansen, 2019; Gobet et al., 2001, for reviews), chunking is mostly ignored in L2-learning measures.

Inspired by these theoretical considerations about chunking in language learning and processing, we present initial results from a novel variation on the elicited imitation paradigm previously used in L2 learning research (e.g., Ellis, 2005; Erlam, 2006; Naiman, 1974; Wu & Ortega, 2013; for review and meta-analysis, see Yan, Maeda, Lv, & Ginther, 2016). Our language measure aims to capture L2 proficiency in the context of more naturalistic stimuli than what is typically used in L2 studies.

We first further discuss the concept of language proficiency, and then examine previous research on methods for evaluating learners. Building on this previous research, our utterance recall test asks L2 learners to listen to naturalistic utterances in the target language and write down what they can remember of the utterance after they finish listening. This test can easily be constructed from the countless sources of authentic audio on the internet. For example, in our study, we collected audio from Netflix[5] television programs. Where captions are available, the entire test creation process can be automated by selecting target utterances of the desired length in the caption file, and automatically extracting the audio using the time stamps in the captions. Importantly, the test can be administered in just a few minutes, while still providing considerable information about a learner's proficiency (as we discuss further below).

In a study with 90 participants, we show that this measure is strongly correlated with students' comprehension (measured by having students provide open-ended, comprehension-focused translations of a separate set of heard utterances). Indeed, our measure is a better predictor of comprehension-based listening-translation ability than a test derived from a standardized multiple-choice comprehension assessment. With this task, we hope to offer a better language proficiency measure for educators and researchers with interests in L2 learning.

Measuring Second Language Proficiency

L2 proficiency has traditionally been measured through comprehensive tests involving multiple-choice or true-false questions (e.g., JLPT and NSE). One downside of such tests is that they take a long time to create and are prone to certain types of confounds. In particular, multiple-choice tests contain information in the question and response options that can help students to strategically choose the correct answer (e.g., Daneman & Hannon, 2001; Katz, Lautenshalger, Blackburn, & Harris, 1990). For example, a study comparing students who read a comprehension passage before answering multiple-choice questions about that passage with students who just read the questions and response options, found that there was no significant difference in the number of correct answers (Bernhardt, 1983). While it may be possible to construct multiple-choice tests that avoid this issue, these tests require significant effort to build because not only does the test creator need to identify the passage, select a question, and create multiple plausible response options, while ensuring that the correct answer cannot be identified using the question and foils alone. Some initial work has explored the possibility of automatically generating multiple-choice questions (Papasalouros, Kanaris & Kotis, 2008), but this approach requires a deep semantic understanding of the text being evaluated. This is especially difficult in non-English languages where data for computer-based topic modeling is scarcer, and fewer researchers are exploring these challenges. This combination of issues makes multiple-choice tests less than ideal for assessing L2 learners.

Another common assessment of language ability is the cloze test (Taylor, 1953). In this task, learners read a passage where some words are deleted, and are asked to fill in the missing words. For example, the learner may see the sentence “I drove to the ____ to buy some eggs” and need to fill in the word “store.” There are many variations in how words can be deleted. For example, some tests randomly select words, others use constant-frequency deletions (e.g., every fifth word), while others may utilize human-selected words based on the specific aspects of grammar or vocabulary being tested. However, it is difficult to compare proficiency levels across different tests because the difficulty of these tests is strongly affected by the specific method used to delete words (Bachman, 1985). Moreover, although these tests have been shown to be reliable measures of reading ability, this deletion paradigm is hard to adapt to audio contexts. It is difficult to cleanly remove words from fluent speech because of coarticulation between words, and often results in rather jarring speech samples.

Many L2 proficiency tests moreover tend to break down language skills into separate components such as grammar and vocabulary (e.g., Alderson & Banerjee, 2002; JLPT; NSE). However, recent research has shown that our use of language does not easily fit into separable grammar and vocabulary components (e.g., Christiansen & Arnon, 2017; Conklin & Schmitt, 2012; Culicover, Jackendoff, & Audring, 2017). For example, consider the garden path sentence “While Anna dressed the baby spit up on the bed.” Most listeners will hear the two phrases: “while Anna dressed the baby” and “the baby spit up on the bed.” To the degree that grammatical regularities shapes our processing of language, we would expect that listeners would resolve the sentence to the grammatical interpretation of “[while Anna dressed] [the baby spit up on the bed].” However, most people do not repair their misunderstandings of garden path sentences like this and instead maintain representations of Anna dressing the baby and the baby spitting up on the bed (Ferreira & Henderson, 1991). Thus, our propensity to group together frequently co-occurring words may better describe how we process this type of sentence than grammatical rules (e.g., Arnon & Christiansen, 2017; Christiansen & Chater, 2016).

Research summarized by Christiansen and Chater (2016) suggest that when we listen to language, we must rapidly process and chunk incoming language, because our memory for auditory information is very limited and is constantly being overwritten by new incoming input. This can be easily demonstrated by listening to an audio segment from a foreign language and attempting to recall as much of that audio as possible. Most people can only accurately recall a few hundred milliseconds of the audio because our short-term memory is limited to only a few elements, ranging from 7 ± 2 (Miller, 1956) to 4 ± 1 (Cowan, 2000). For example, we might have difficulty recalling even the 10 digits of a US phone number when spoken aloud (which is why we might try to group the 10 digits into three chunks, corresponding to the area code, exchange code, and line number: xxx-yyy-zzzz). The limits of short-term memory suggest that a fundamental aspect of language proficiency is our ability to quickly process incoming information and chunk it together into meaningful units. That is, rapidly chunking language input appears to be central to language proficiency. This perspective has been substantiated by computational modeling work (McCauley & Christiansen, 2019), which also demonstrated that chunking can provide a common basis for both comprehension and production (Chater, McCauley, & Christiansen, 2016) and even provide insights into differences between first language (L1) and L2 learners (McCauley & Christiansen, 2017). Thus, even though the chunking perspective primarily has focused on L1 acquisition, the same constraints apply to processing more broadly, suggesting that chunking may be essential to L2 proficiency as well.

Standard tasks used to test L2 proficiency, such as multiple-choice and cloze measures, do not tap into chunking ability. While less commonly used, a recall paradigm has more promise as an L2 assessment that targets the key role of chunking in language learning and use. In the memory literature, serial recall of lists of items (words, letters, digits) have been used extensively to measure the effect of chunking on memory abilities (e.g., Jones & Macken, 2015). Variations of this task have been used to test general statistical learning (e.g., Isbilen et al., 2017, in press)—the ability to learn distributional patterns of co-occurrence in language and other aspects of cognition (see Frost, Armstrong, & Christiansen, 2019; Rebuschat & Williams, 2012, for reviews). Under the guise of "sentence imitation," recall of whole sentences has long been used to assess L1 acquisition in children (e.g., Frizelle, O'Neill, & Bishop, 2017;

Slobin & Welsh, 1967). In this imitation task, participants listen to an utterance and repeat it out aloud, providing a useful measure of language processing ability.

The elicited imitation paradigm has also been used previously in L2 learning (e.g., Bowden, 2016; Ellis, 2005; Hamayan, Saegert, & Larudee, 1977; Suzuki & DeKeyser, 2015; Wu & Ortega, 2013; see Yan et al., 2016, for a review). Erlam (2006) suggests that elicited imitation is reconstructive. That is, learners must use long-term knowledge of the language in order to complete the task because short-term memory is too limited to retain information about an entire utterance (see also Klem et al., 2015). High-proficiency learners also correct grammatical mistakes in recalled utterances, suggesting that the meaning of the utterance is remembered rather than the words ad verbatim (Hamayan et al., 1977). Another study used a recall task designed to test comprehension, asking students to read a passage and then immediately write down as much information from the passage as they could remember (Bernhardt, 1983). An instructor identifies “idea units” from the original passage and students’ responses are coded for the number of idea units that are included. For example, the sentence “The professor does research on spiders” might have units for “the professor” and “does research.” This test gives a comprehensive picture of comprehension, but needs a trained human to score and adds an element of subjectivity into the measure. For example, different researchers might choose larger or smaller idea units.

We use the prior work on sentence imitation as a starting point for our recall measure, viewing it as a natural-language chunking task (Chater et al., 2016; Christiansen, 2019; Christiansen & Chater, 2016). Previous research using elicited imitation has suggested that this task provides a useful measure of general L2 proficiency (e.g., Bowden, 2016; Ortega, 2000; see, Yan et al., 2016, for a review). Moreover, since the landmark study of Ellis (2005), elicited imitation has also been construed as a potentially useful task for capturing implicit aspects of L2 learning (e.g., Bowles, 2011; Erlam, 2006)—though this is currently a matter of some debate (e.g., see Suzuki & DeKeyser, 2015, for a different perspective). Our study is not intended to address this debate but, rather, aims to provide a novel measure of real-time L2 skills, while also allowing for automatic scoring of the results. Specifically, we aim to test 1) the feasibility of using naturalistic stimuli that better reflect the kind of conversational input that learners may experience in the real world, and 2) automatic methods for scoring elicited imitation and translation abilities.

Table 1 Participant language background information

	Min	Max	Mean	*SD*
Self-reported Spanish proficiency	1	7	3.26	1.35
Years of high school level Spanish	0	6	2.77	1.70
Years of college level Spanish	0	3	0.27	0.63
Years learning independently	0	21	1.27	3.73
Years in a Spanish speaking country	0	3	0.07	0.36

Method

To evaluate our approach to L2 learner proficiency, we asked participants to complete a battery of tasks that, in addition to our novel utterance recall task, included a test of learners' general comprehension skill (using an open-ended, comprehension-based translation test), and a multiple-choice comprehension task based on a standardized test (a subset of the NSE). Following open science practices, our materials are available from https://osf.io/guem7/.

Participants

Data was collected online through Qualtrics survey software and 100 participants were recruited through a university research participation system. All participants had some experience with Spanish language learning, including high-school or college-level Spanish classes or independent learning (see Table 1 for summary data). Because we aimed to gauge the effectiveness of our measure across a broad range of skill levels, we allowed participants of any level to take part in the study. Spanish was used as the foreign language, because it provided a large and diverse pool of participants. Data from 90 participants were included in the final analysis (three participants were excluded from analysis because they had technical issues during the study and seven others because they were not native speakers of English).

Tasks and Materials

We considered two factors when designing the study. First, we aimed to design the naturalistic utterance recall measure such that a similar test could easily (either automatically or with minimal researcher input) be created for other languages. Given studies suggesting that working memory abilities are affected by the nature of individual languages (Amici et al., 2019; MacDonald & Christiansen, 2002), including in L2 contexts (Van den Noort, Bosch, & Hugdahl, 2006), we administered recall tasks in both L2 (Spanish) and L1 (English).

We therefore expected that Spanish L2 recall would be a better predictor of proficiency (measured via our translation test) than English L1 recall. Second, we included a shortened version of the NSE to provide a more conventional baseline for language-learning proficiency.

Spanish Recall Task

To ensure that the test could be easily constructed (and that variants could easily be created), we used online television programs as an audio source. Platforms such as Netflix and YouTube have countless hours of foreign language video. Many of these videos are captioned and, if not, captions can easily be added to short segments of the video. Furthermore, the audio is likely to be more similar to what a learner may encounter in real-world interactions than stimuli produced by a speech synthesizer. When captions are included with the video, it is easy to automatically select utterances for the test (using a combination of utterance length and the vocabulary in the caption) and extract the audio based on the caption time stamps. Audio clips of between 5.5 and 6.5 seconds were selected based on informal pilot studies[6] that showed this duration was just beyond the maximum length that most native speakers are able to remember. To determine whether a short test would be able to measure L2 proficiency, participants were asked to recall five different Spanish utterances. Thus, a total of five clips were used with word counts ranging from 21 to 31 and an average of 24.8 words each (*SD*: 4.27). Each clip was produced by a different speaker (40% female). Complete transcriptions of the clips can be found in Appendix S1.

To supplement our Spanish recall task, we included a debriefing question asking participants whether or not they translated the Spanish utterances into English during the recall task. This allowed us to assess any potential differences between participants who focused on remembering using only Spanish and those who completed the task by translating heard utterances into English and then back to Spanish.

English Recall Task

To factor out the potential contribution of basic memory abilities to the Spanish recall task, we also administered an English recall task to assess whether variation in the Spanish recall task could be explained by general differences in auditory chunking ability rather than L2 proficiency. Participants were asked to recall utterances from five different English clips that had word counts ranging from 16 to 33, and an average of 22.4 words (*SD*: 6.35). Each clip was produced by a different speaker (40% female). Complete transcriptions of the clips can be found in Appendix S2.

Multiple-Choice Task

To provide a standardized measure of L2 comprehension ability, we included a set of eighteen multiple-choice questions from the NSE (https://www.nationalspanishexam.org/) offered by the American Association of Teachers of Spanish and Portuguese. This test was chosen because it is used in many classrooms in the United States to measure Spanish proficiency[7] and past exams are freely available. Three questions were randomly drawn from each of the six difficulty levels of the exam (total of 18 questions) to create a more standard proficiency assessment. We acknowledge that choosing a random subset of questions could compromise the integrity of this test. Nonetheless, insofar as the test has a reasonable degree of internal consistency, this should not invalidate the results obtained here, though care should be taken in interpreting the results. Moreover, we primarily included this multiple-choice task for reference, to be compared with our Spanish recall task.

Translation Task

As our dependent variable of L2 comprehension proficiency against which to evaluate our recall measure, we needed a robust and ecologically valid task. Therefore, we designed a translation task to attain the most complete picture of a learner's ability to comprehend language in real time. Our translation task is based on the Listening Summary Translation Exam developed by Scott, Stansfield, and Kenyon (1996). In their task, learners listen to entire conversations in another language and then summarize those conversations in English. This task was developed for FBI employees who would need to summarize conversations as part of their work. We modified the test to better reflect conversational proficiency and to better accommodate low-skill learners. During real-life language interactions, a learner is faced with the challenge of quickly comprehending what a speaker says in order to provide an appropriate response in a timely matter (within less than a quarter of a second; see Levinson, 2016, for a review). Although a learner can always request clarifications during typical real-world interactions, to accommodate the fast-paced flow of a normal conversation, learners will need to comprehend the majority of what they hear in the first place. Our translation task was designed to mimic this type of scenario. The learner hears a short audio clip and is asked to translate what was heard. Participants heard a total of eighteen separate clips ranging from 11 to 18 words (avg. 15.1, *SD*: 2.04). The clips were produced by six different speakers (83% female). Complete transcriptions of the clips can be found in Appendix S3.

By keeping these clips relatively short, we minimize potential memory confounds that have been identified as a pitfall of this testing method (Alderson & Banerjee, 2002). Assuming that the learner understood what was heard, they should be able to express this information in a translation even if the translation deviates from the literal meaning of the heard utterance. This type of open-ended translation allows us to estimate how much of a given sentence a participant is able to understand.

In sum, our experiment included Spanish and English recall tests, a set of comprehension multiple-choice questions from a standardized test, and an auditory translation task. We view the open-ended audio translation task as an ecologically valid assessment of a learner's L2 comprehension ability (our dependent variable). In our analyses, we determine which measures (recall or multiple choice) best predict a learner's L2 proficiency as assessed by the translation task.

Procedure

Our study was conducted remotely, using the participants' own computer. This means that we did not have complete control over the testing environment. While it is possible that this may affect test accuracy for individual participants, this would only add noise to the data, which would work against our hypotheses. We would therefore expect lab-based replications of our study to show even stronger effects.

Participants first read a consent form and agreed to take part in the study. They completed a short exercise to ensure that the audio was working on their computer: they listened to an audio clip and wrote down the word "communicate." Participants then provided some basic demographic information about their native language, proficiency in Spanish, and so on (see Appendix S4 for the complete questionnaire). Next, they completed the English recall task (as a baseline for auditory recall ability), the Spanish recall task, followed by the translation task.

For both the translation and recall tasks, participants were instructed to click a button to begin each audio segment. While the audio played, the answer boxes were hidden so that participants could not recall or translate the segments word-by-word. Once the audio finished playing in the recall tasks, a single text entry box appeared along with text instructions ("Write what you just heard below."). In the translation task, participants were asked to write down an English translation of the Spanish clip in the text box. Each audio clip was presented only once, after which the participant entered their response. The recall task can thus be viewed as a transcription task, where participants write down as much of the verbatim utterance as possible. Participants completed a practice recall trial before both the English and Spanish recall tasks. The instructions for the Spanish recall task asked the participants not to translate the audio clips into English. In the translation task, participants were instructed to translate as much as they could, while noting that they may not be able to translate everything they heard. In total, participants completed five English transcriptions, five Spanish transcriptions, and 18 translations from Spanish to English. Following the Spanish recall task, participants were asked if they were mentally translating during the task.[8]

Next, participants completed the multiple-choice task. Each question had four response options and participants were instructed to choose the correct answer. A one-sentence context along with the question and response options was displayed to participants. Once participants were ready, they clicked a button to begin the audio clip. These audio clips were 20–40 seconds long. In order to mimic typical testing conditions, participants did not have to wait till the end of the audio clip before making their response but could do so at any time after the beginning of the trial. However, participants were required to choose an answer before continuing to the next trial.

Finally, participants were asked if they had any questions about the study and thanked for their time.

Analysis

A key issue with both recall and translation tasks is that scoring the responses not only can be labor intensive but also introduce subjective elements into the results. We therefore adapted two algorithm-based scoring methods from the natural language processing literature: word error rate (WER; Evermann, 1999) and semantic similarity (Han, Kashyap, Finin, Mayfield, & Weese, 2013). As described further below, we use WER to score recall performance and semantic similarity to assess translation accuracy.

Word Error Rate

Previous L2 studies using elicited imitation have involved human raters, who either score a recalled item in a binary fashion (correct vs. incorrect; e.g., Ellis, 2005; Erlam, 2006) or on an ordinal scale (e.g., 0–4, 4 = perfect recall; Bowden, 2016; Ortega, 2000). By contrast, WER provides a more fine-grained, automatic measure of recall. WER captures the difference between a response item and the target (the so-called "gold standard"), by computing how many changes need to be made to the former to produce the latter. Specifically, it is calculated as: deletions + insertions + substitutions/words in the target (Evermann, 1999). A deletion is a missing word when compared to the target (e.g., "The cat runs," target: "The black cat runs"), an insertion is an additional word when compared to the target (e.g., "The brown cat runs," target: "The cat runs"), and a substitution is a changed word (e.g., "The brown cat runs," target: "The black cat runs").

The WER metric has been used extensively to evaluate the quality of machine translation and speech recognition systems[9] (e.g., Evermann, 1999). Here, we use WER to automatically score performance on the Spanish and English recall tasks (as described further below, we also used WER as part of one of our two measures of translation performance). We used the transcriptions from the video captions as the targets to be compared with the response items. These were verified by a native Spanish speaker for accuracy. To facilitate comparisons with semantic scores, our reported statistics and plots use 1-WER rather than the WER score directly. Furthermore, the scores are summed across all trials. Thus, for Spanish and English recall, the possible scores range from 0 to 5, with 5 being a perfect score.

Semantic Similarity

When translations are used as a measure of L2 proficiency, scoring responses typically involves the use of human raters, which can introduce subjectivity into the evaluation process (e.g., Scott et al., 1996). To provide an automatic and objective assessment of translation quality, we adopted a standard measure of semantic similarity (Han et al., 2013) from the machine learning literature. Semantic similarity provides a measure of how similar the overall meaning of two different texts are to one another. Because we were more interested in overall comprehension than exact wording, semantic similarity between a participant's translation and the official caption provides a fully automatic way to assess the participant's ability to capture the intended meaning of the utterance (i.e., this measure does not require any human raters to evaluate the translations). We, thus, used semantic similarity to determine the match between participant-generated translations and the English captions.

The semantic similarity scores were calculated using the API (application programming interface) provided by Han et al. (2013). Their method combines output from a thesaurus (WordNet; Miller, 1995) and corpus analysis (Web corpus from Stanford WebBase project; Stanford, 2001) to calculate semantic similarity. Semantic similarity is calculated with a model that uses features such as word similarity and overlapping n-grams (sequences of n consecutive words). Word similarities are calculated by identifying words that appear in similar contexts and overlapping n-grams are created by aligning words in the input phrase with the highest similarity words in the target sentence. This approach means that words do not need to exactly match the words in the target translation but that the relationship between words is also considered. Thus, semantic similarity provides a measure of comprehension as reflected in the translation, rather than English writing skills.

Human Ratings of Translations

As a second, more standard assessment of translation performance, we asked human raters to evaluate the translations. Because the goal of the translation task was to assess learners' general comprehension of the heard utterances, rather than their ability to translate the utterances verbatim, we used human raters to evaluate how well a participant's translation captured the intended meaning of the utterance. Two bilingual speakers of English and Spanish transformed each participant's translation into a gold standard translation while minimizing the changes made. For example, for the Spanish phrase "es necesario decirlo de frente" ("I have to be very direct"), a learner might produce "it is necessary to say" which was transformed into "it is necessary to speak directly." This way we ensure that the translation scores (derived via WER) accurately reflect how much the participant understood the intended meaning of an utterance, while ignoring particulars that do not directly influence meaning.

The raters were told to create a correct translation out of each learner's translation by adding and removing as few words as possible. Importantly, the raters were blind to our research questions. Discrepancies between raters were resolved in two stages. First, the guidelines were reviewed, and both of the raters were shown the two translations and asked to produce a new translation that most closely followed the original guidelines. Following the first round, 421 of the 1,746 (24%) translations were reevaluated by the raters. In the second round, remaining discrepancies were discussed, and a final gold standard translation was agreed upon by the raters. In total, 87 of the 1,746 (5%) translations were discussed in the second round. Because raters produced translations rather than a numerical rating, we do not report additional rater agreement metrics.

We used WER to quantify participant translation performance relative to the gold standard provided by the human rater. To facilitate comparison with our semantic similarity measure, we again report scores as 1-WER and sum scores across trials, with scores ranging from 0 to 18 (18 being a perfect score).

Table 2 Descriptive statistics for proficiency measures

Measure	Min	Max	Mean	*SD*
Translation (semantic)	0	10.73	3.69	2.47
Translation (1-WER)	0	16.50	3.81	3.47
Spanish recall (1-WER)	0.03	3.26	0.72	0.61
English recall (1-WER)	0.18	4.31	2.92	0.79
Multiple-choice	3	17	11.34	3.11

Table 3 Correlations between proficiency measures

	Translation (semantic)	Translation (1-WER)	Spanish recall (1-WER)	English recall (1-WER)	Multiple-choice	Self-report
Translation (semantic)	—	0.909^{***}	0.766^{***}	0.412^{***}	0.548^{***}	0.620^{***}
Translation (1-WER)		—	0.895^{***}	0.436^{***}	0.586^{***}	0.680^{***}
Spanish recall (1-WER)			—	0.394^{***}	0.591^{***}	0.678^{***}
English recall (1-WER)				—	0.483^{***}	0.374^{***}
Multiple-choice					—	0.454^{***}
Self-report						—

Note. $^{*}p < .05$, $^{**}p < .01$, and $^{***}p < .001$ after Bonferroni correction.

Results

Following open science practices, all participant data are available at https://osf.io/guem7/. As expected, we observed considerable variability across the different tasks, as indicated by the descriptive statistics in Table 2. We observed substantial correlations between performance in the various tasks, as shown in Table 3. The two variations of our dependent measure, semantic similarity and WER, are strongly correlated with one another ($r = 0.909, p < .001$), suggesting that our automatic and objective measure of translation performance, semantic similarity, may provide a useful, easy-to-use substitute for more labor-intensive human-based translation ratings. Importantly, given our hypothesis, Figure 1 illustrates the strong correlation between semantic similarity scores and Spanish recall ($r = 0.766$, $p < .001$): the better the recall of Spanish utterances, the

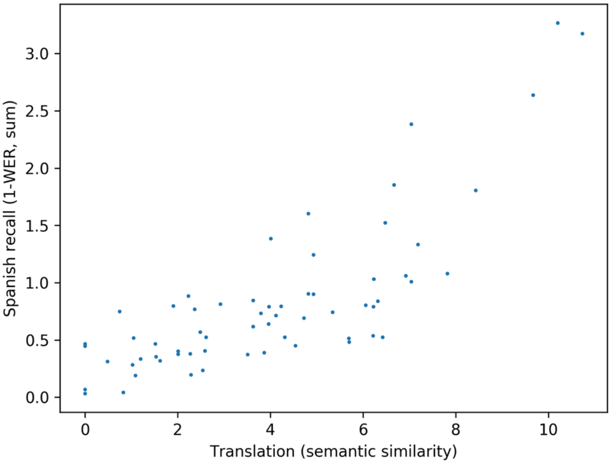

Figure 1 A strong correlation between Spanish utterance recall and the semantic similarity score showing that our Spanish recall measure predicts translation performance when scored automatically using semantic similarity. [Color figure can be viewed at wileyonlinelibrary.com]

better the participant is at capturing the semantic content when translating from Spanish into English. In a similar vein, Figure 2 shows that there is also a strong correlation between Spanish recall scores and Spanish to English translation scores using WER translation scores ($r = 0.895$, $p < .001$). By contrast, Figure 3 shows that there is considerably weaker, but still moderate correlation between Spanish recall and the standard measure using multiple-choice scores ($r = 0.591$, $p < .001$). Interestingly, we obtained a reasonably strong correlation between Spanish recall scores and self-reported proficiency ($r = 0.678$, $p < .001$), as illustrated in Figure 4. Self-reported proficiency is also strongly correlated with Spanish to English semantic similarity translation scores ($r = 0.620, p < .001$), suggesting that both of these automatic measures align with participants' self-perception of their language skills.

A partial correlation analysis was conducted to assess whether general auditory memory ability (as measured by the English recall task) played a significant role in a participant's Spanish recall ability. If the Spanish recall measure is significantly affected by general memory, this would limit the effectiveness of recall as an L2 proficiency measure. However, English recall accounted for only a small part of the overall correlation ($r_{part} = 0.187$) in a model predicting semantic similarity translation score using Spanish recall and English recall. This supports our hypothesis that the Spanish recall task measures chunking

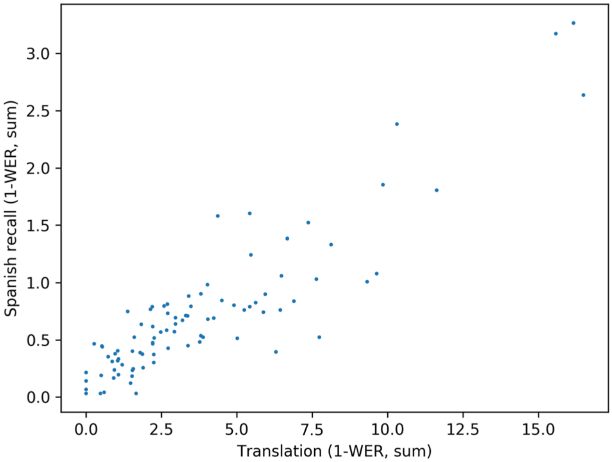

Figure 2 A strong correlation between Spanish utterance recall and translation accuracy indicating that our recall measure is also a good predictor of translation ability measured via human raters. [Color figure can be viewed at wileyonlinelibrary.com]

ability relevant for processing Spanish, largely independent of participants' native language chunking ability in English. Indeed, the English recall score was only weakly correlated with both the semantic similarity ($r = 0.412, p < .001$) and the WER ($r = 0.436, p < .001$) translation scores. Furthermore, English recall was also only weakly correlated with multiple-choice performance ($r = 0.483, p < .001$) and self-reported Spanish proficiency ($r = 0.374, p < .001$).

Another partial correlation was calculated to assess whether clip length had a significant impact on the number of words recalled. Although the duration of the audio clips differed by at most 1 second, it is important to determine whether or not our recall measure is overly sensitive to variations in audio clip length. The partial correlation was found to be small ($r_{part} = 0.113$) in a model predicting the semantic similarity translation score using recall for an individual audio clip and the duration of that audio clip in seconds. This suggests that small differences in audio clip duration may not affect the test outcome using the recall measure.

High- and Low-Proficiency Learners

To assess whether the recall measure is better suited for low- or high-proficiency learners, a median split was performed on the translation semantic similarity scores, and the lower and upper quartiles were analyzed independently. We

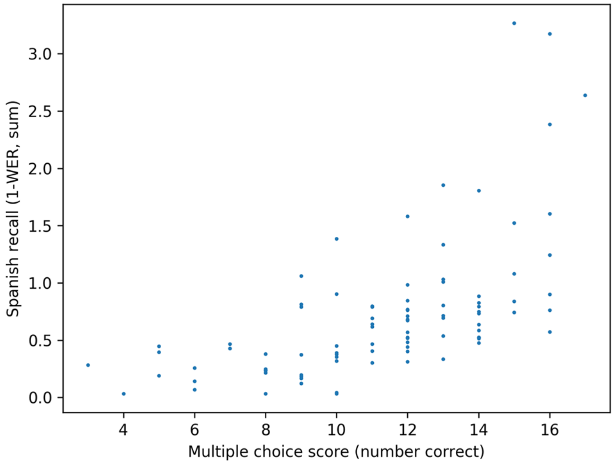

Figure 3 A moderately strong correlation between Spanish utterance recall and performance on the standard multiple-choice task suggesting that our recall measure captures a decent amount of the variance in the multiple-choice test. [Color figure can be viewed at wileyonlinelibrary.com]

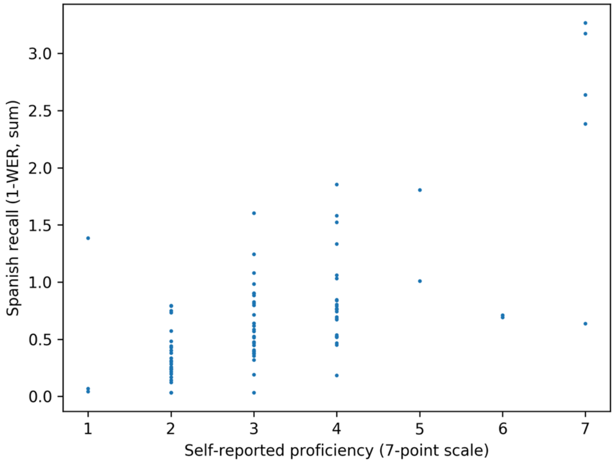

Figure 4 A moderately strong correlation between Spanish utterance recall and the learner's self-reported Spanish proficiency indicating that our recall task captures some of the learner's perception of their own second language skills. [Color figure can be viewed at wileyonlinelibrary.com]

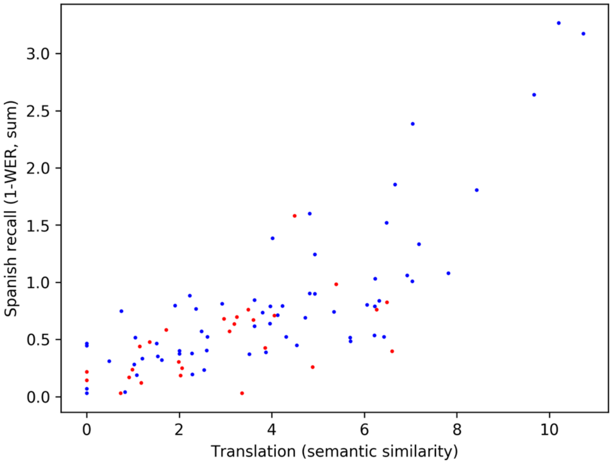

Figure 5 A replot of the correlation between Spanish utterance recall and the translation semantic similarity score, with red points denoting participants who mentally translated during the task and blue points the participants who did not. Mental translation leads to poorer recall and lower translation scores. [Color figure can be viewed at wileyonlinelibrary.com]

found that correlations were greater in the upper quartiles ($r = 0.704, p < .001$) compared to the lower quartiles ($r = 0.391, p = .007$). This suggests that the test may be less sensitive to skill differences among beginners compared to distinguishing between high- and low-proficiency learners, or identifying skill differences between high-proficiency learners. However, it is also possible that the somewhat weaker correlation for the bottom quartile may be due to a more constricted range of translation proficiency among these individuals compared to those in the upper quartiles.

Mentally Translating During the Spanish Recall Task

Although participants were instructed not to mentally translate into English during the Spanish recall task, we were concerned that some participants may nonetheless use translation during recall. After completing the task, we therefore asked participants whether they had mentally translated or not. A total of 27 participants reported translating during the task. As illustrated by Figure 5, participants who reported that they translated during recall tended to have lower translation scores, suggesting that mental translation is mostly used by lower proficiency learners. The correlation between translation semantic similarity scores and Spanish recall scores was substantially lower for those participants who translated ($r = 0.578, p < .001$) than for those participants who did not ($r = 0.788, p < .001$).

Discussion

In this study, we have provided initial proof of concept that recall of naturalistic audio clips may provide a useful and easy-to-administer measure of L2 proficiency. Adding to the ecological validity of our study, the stimuli for both the recall and translations tasks were produced by different speakers of both genders. More generally, our approach is based on recent work in L1 acquisition, emphasizing the importance of rapid processing of linguistic input in the here-and-now by way of memory-based chunking processes (e.g., Christiansen & Chater, 2016). Accordingly, we used audio clips taken from conversational exchanges in TV shows so that our measure of L2 proficiency would better approximate the online processing requirements of real-world interactions. Methodologically, our study moreover adds to the literature on elicited imitation as a measure of L2 proficiency (e.g., Bowden, 2016; Ellis, 2005; Erlam, 2006; Ortega, 2000) by demonstrating that an objective, automatic computational procedure can be used to score recall performance (as well as translation ability). Below, we situate our findings in the broader context of L2 research, discuss current limitations, and avenues for further research.

Improvement Over Multiple-Choice

Our findings dovetail with previous studies using elicited imitation, suggesting that utterance recall provides an efficient measure of L2 proficiency (see also the meta-analysis by Yan et al., 2016, confirming that such tasks have a strong ability to discriminate between L2 learners of different levels of proficiency). Our comparison with the shortened multiple-choice task (based on the standardized NSE test) suggests that such tasks may provide relatively weak measures of proficiency. In our study, self-reported proficiency level was more highly correlated with participants' translation performance ($r = 0.620$) than the multiple-choice score ($r = 0.454$). We acknowledge, though, that the shortened version of the task may have affected its integrity but insofar as the overall task is internally consistent, our results should still be interpretable (while keeping this caveat in mind). Indeed, we observed similar results during our pilot studies using a different subset of multiple-choice questions from the same test.

As previously discussed, there are a number of drawbacks to using multiple-choice questions (e.g., Bernhardt, 1983; Daneman & Hannon, 2001; Katz et al., 1990), especially in the context of online learning. In our analysis, we found that the information that can be learned about the student is sparse when using multiple-choice questions. We can only gain a single binary correctness signal for each question answered: the student answered the question correctly or they did not (a similar issue pertains to the binary scoring used in some elicited imitation studies—e.g., Ellis, 2005; Erlam, 2006—which appears to be less sensitive to differences in levels of proficiency than ordinal scales according to Yan et al., 2016, meta-analysis). This means that many items may be needed to evaluate a student's proficiency, and, combined with the difficulty of developing these tests, this compounds the challenge of creating effective multiple-choice tests. This also fails to utilize the student's time effectively. The student must spend significantly more time listening and responding to questions when multiple-choice questions are used because they not only need to listen to the prompt, but also the question and each response option. Although multiple-choice questions continue to be used in many proficiency tests (e.g., JLPT and NSE), we suggest that our utterance recall measure might allow for better use of students' and teachers' time, and increase the accuracy of the tests.

Implications for Design of Language Proficiency Tests

We have shown that a simple utterance recall test based on readily available television programs and subtitles can be used to design accurate tests of L2 learning proficiency in more naturalistic contexts. While some care needs to be taken to avoid particularly noisy clips, in general audio tracks from videos are designed to be understood, so most of the audio can be used.

In our utterance recall test, we used audio clips that pilot data had suggested that most participants would not be able to fully recall. No participant was able to recall every word. While this makes the task more difficult for participants, it helps to avoid ceiling effects, which we saw some evidence of in the multiple-choice test. Previous work has shown that very short utterances can be recalled with rote memorization (Erlam, 2006), so we recommend choosing longer target utterances when employing recall measures like the one we have used here. Moreover, longer utterance may also be more representative of the language that learners will encounter in the real world.

The meta-analysis of elicited imitation studies by Yan et al. (2016) showed that several factors influence the sensitivity of this task, including sentence length, scoring method, and construct (e.g., global proficiency, morphosyntactic). In future work, it will be important to evaluate how many recall utterances may be needed, and the optimal length of clips. In the present study, the length of an audio clip did not affect the number of words recalled, suggesting that scores may be comparable across clips, provided they are long enough to avoid ceiling effects. In contrast to previous work using elicited imitation (e.g., Ortega, 2000), we did not manipulate lexical difficulty and syntactic complexity directly in this study. We therefore encourage future users of our method to explore how these factors might affect utterance recall, which may improve accuracy even further. Furthermore, the recall test used here does not provide an absolute measure of proficiency, only relative proficiency between learners. Subsequent work could investigate whether recall scores can be used to create an absolute measure of proficiency to improve comparison across studies (e.g., by manipulating the syntactic complexity of the stimuli).

What L2 Knowledge Does Naturalistic Utterance Recall Measure?

Recent work in psycholinguistics has highlighted the key importance of processing in the here-and-now, for dealing both with the onslaught of linguistic input given memory limitations (Christiansen & Chater, 2016) and the rapid pace of turn-taking during normal conversation (Levinson, 2016). Building on these considerations, our use of naturalistic audio clips was intended to measure L2 proficiency as relevant for language processing in real-life conversations. Thus, we hypothesize that our utterance recall task provides a general measure of auditory L2 proficiency, including knowledge of vocabulary (known words are easier to process and recall), grammatical patterning (known constructions are easier to comprehend and recollect), as well as phonology and prosody (fast real-time processing requires familiarity with phonological categories and prosodic patterns). Indeed, the meta-analysis by Yan et al. (2016) found elicited imitation was more sensitive when used as a global construct (e.g., Ortega, 2000, and here) compared to when used to measure more narrow constructs such as morphosyntax (e.g., Bowles, 2011; Erlam, 2006).

As a measure of L2 proficiency, our recall task is likely to rely primarily on implicit L2 knowledge (e.g., Bowles, 2011; Ellis, 2005; Erlam, 2006), though automatized explicit knowledge may also play a role (e.g., Suzuki, 2017; Suzuki & DeKeyser, 2015). However, our study was not designed to determine the relative contributions of explicit vs. implicit knowledge to the performance on our utterance recall task. We leave it for future studies to investigate this relationship.

However, it is important to acknowledge that although our utterance recall task appears to provide a good index of listening comprehension (and translation skills), it would likely need to be adapted somewhat to capture oral and written L2 proficiency. For example, the stimuli to be recalled could be presented in written form to gauge the understanding of printed material. And recall of auditory stimuli could be done verbally to assess oral L2 skills (though this would not capture free-form conversational language production). Thus, we believe that it might be possible to extend the current recall-based approach to measure additional aspects of L2 proficiency.

Conclusion

In this study, we have presented a new variation on the elicited imitation task, aiming to measure L2 listening proficiency relevant to real-time language processing. We have also shown that our task allows for objective computer-based scoring. Despite its shortcomings, we hope that this easy-to-administer and easy-to-score test may help accelerate the development of effective foreign language learning methodologies that consider a learner's online processing ability instead of grammar and vocabulary tests, which fail to provide a complete picture of language proficiency. Furthermore, we have shown that naturalistic native-speaker materials can be used not only as learning resources, but also as assessment tools. By using these resources in assessments, learning is evaluated with tasks that more closely resemble the real-life situations where learners will eventually use their language skills.

Final revised version accepted 3 February 2020

Notes

[1] According to: http://www.jlpt.jp/e/guideline/testsections.html

[2] According to: https://www.nationalspanishexam.org/index.php/exam-administration/exam-length

[3] Text from dialogues in *Genki I* and *Genki II* (http://genki.japantimes.co.jp/index_en) was tokenized and unique tokens were counted.

[4] Text from the subtitles of *Tokyo Diner* (https://www.netflix.com/title/80113037) was tokenized and unique tokens were counted.

[5] https://www.netflix.com/

[6] In total, 70 participants took part in our pilot studies, the goal of which was to improve the recall tests, resulting in multiple different changes to the tasks over time. Because these studies were entirely exploratory, we do not provide any analysis of these data.

[7] According to the NSE website, "The National Spanish Examinations are the most widely used tests of Spanish in the United States. In the spring of 2019, a total of 152,069 students registered for the National Spanish Examinations." (https://www.nationalspanishexam.org/index.php/about-us/what-is-nse accessed December 17, 2019).

[8] Specifically, we asked the participants the following question: “In the task you just completed, where you wrote down all of the Spanish you could remember, did you mentally translate the phrases into English before writing down the Spanish?”

[9] In addition to the WER analysis, an analysis using BilLingual Evaluation Understudy (BLEU) was also conducted (Papineni, Roukos, Ward & Zhu, 2002). However, the BLEU results were very similar to the WER results, so we only present results of the WER analysis.

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the components of the research methods needed to reproduce the reported procedure. All materials that the authors have used and have the right to share are available at https://osf.io/guem7/. All proprietary materials have been precisely identified in the manuscript.

References

Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching*, *35*, 79–113. https://doi.org/10.1017/S0261444802001751

Amici, F., Sánchez-Amaro, A., Sebastián-Enesco, C., Cacchione, T., Allritz, M., Salazar-Bonet, J., & Rossano, F. (2019). The word order of languages predicts native speakers’ working memory. *Scientific Reports*, *9*, 1124. https://doi.org/10.1038/s41598-018-37654-9

Arnon, I., & Christiansen, M. H. (2017). The role of multiword building blocks in explaining L1-L2 differences. *Topics in Cognitive Science*, *9*, 621–636. https://doi.org/10.1111/tops.12271

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, *19*(3), 535–556. https://doi.org/10.2307/3586277

Bernhardt, E. B. (1983). Testing foreign language reading comprehension: The immediate recall protocol. *Die Unterrichtspraxis/Teaching German*, *16*, 27–33. https://doi.org/10.2307/3530598

Bowden, H. W. (2016). Assessing second-language oral proficiency for research: The Spanish elicited imitation task. *Studies in Second Language Acquisition*, *38*, 647–675. https://doi.org/10.1017/S0272263115000443

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition*, *33*, 247–271. https://doi.org/10.1017/S0272263110000756

Chater, N., McCauley, S. M., & Christiansen, M. H. (2016). Language as skill: Intertwining comprehension and production. *Journal of Memory and Language*, *89*, 244–254. https://doi.org/10.1016/j.jml.2015.11.004

Christiansen, M. H. (2019). Implicit statistical learning: A tale of two literatures. *Topics in Cognitive Science*, *11*, 468–481. https://doi.org/10.1111/tops.12332

Christiansen M. H., Arnon I. (2017). More than words: The role of multiword sequences in language learning and use. *Topics in Cognitive Science*, *9*, 542–551. https://doi.org/10.1111/tops.12274.

Christiansen, M. H., & Chater, N. (2016). The Now-or-Never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, e62. https://doi.org/10.1017/S0140525X1500031X

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61. https://doi.org/10.1017/S0267190512000074

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral & Brain Sciences*, *24*, 87–185. https://doi.org/10.1017/S0140525X01003922

Culicover, P. W., Jackendoff, R., & Audring, J. (2017). Multiword constructions in the grammar. *Topics in Cognitive Science*, *9*, 552–568. https://doi.org/10.1111/tops.12255

Daneman, M., & Hannon, B. (2001). Using working memory theory to investigate the construct validity of multiple-choice reading comprehension tests such as the SAT. *Journal of Experimental Psychology*, *130*, 208–223. https://doi.org/10.1037/0096-3445.130.2.208

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition*, *27*, 141–172. https://doi.org/10.1017/S0272263105050096

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics*, *27*, 464–491. https://doi.org/10.1093/applin/aml001

Evermann, G. (1999). *Minimum word error rate decoding* (Unpublished master's thesis). Cambridge University, Cambridge, United Kingdom.

Ferreira, F., & Henderson, J. M. (1991). Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, *30*, 725–745.

Frizelle, P., O'Neill, C., & Bishop, D. V. M. (2017). Assessing understanding of relative clauses: A comparison of multiple-choice comprehension versus sentence repetition. *Journal of Child Language*, *44*, 1435–1457. https://doi.org/10.1017/S0305000916000635

Frost, R., Armstrong, B. C., & Christiansen, M. H. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, *145*, 1128–1153. https://doi.org/10.1037/bul0000210

Gobet F, Lane P, Croker S, Cheng P, Jones G, Oliver I, Pine J (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*, 236–243. https://doi.org/10.1016/s1364-6613(00)01662-4.

Hamayan, E., Saegert, J., & Larudee, P. (1977). Elicited imitation in second language learners. *Language and Speech*, *20*, 86–97. https://doi.org/10.1177/002383097702000109

Han, L., Kashyap, A. L., Finin, T., Mayfield, J., & Weese, J. (2013). UMBC_EBIQUITY-CORE: Semantic textual similarity systems. *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, *1*, 44–52.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*, 422–433. https://doi.org/10.1017/S1366728911000678

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (in press). Statistically-induced chunking recall: A memory-based approach to statistical learning. *Cognitive Science*.

Isbilen, E. S., McCauley, S. M., Kidd, E., & Christiansen, M. H. (2017). Testing statistical learning implicitly: A novel chunk-based measure of statistical learning. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 564–569). Austin, TX: Cognitive Science Society.

Jones, G., & Macken, B. (2015). Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. *Cognition*, *144*, 1–13. https://doi.org/10.1016/j.cognition.2015.07.009

Katz, S., Lautenshalger, G. J., Blackburn, A. B., & Harris, F. H. (1990). Answering reading comprehension items without passages on SAT. *Psychological Science*, *1*, 122–127. https://doi.org/10.1111/j.1467-9280.1990.tb00080.x

Klem, M., Melby-Lervåg, M., Hagtvet, B., Lyster, S. A. H., Gustafsson, J. E., & Hulme, C. (2015). Sentence repetition is a measure of children’s language skills rather than working memory limitations. *Developmental Science*, *18*, 146–154. https://doi.org/10.1111/desc.12202

Levinson, S. C. (2016). Turn-taking in human communication—Origins and implications for language processing. *Trends in Cognitive Sciences*, *20*, 6–14. https://doi.org/10.1016/j.tics.2015.10.010

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: A comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35–54. https://doi.org/10.1037/0033-295X.109.1.35

McCauley, S. M., & Christiansen, M. H. (2017). Computational investigations of multiword chunks in language learning. *Topics in Cognitive Science*, *9*, 637–652. https://doi.org/10.1111/tops.12258

McCauley, S. M., & Christiansen, M. H. (2019). Language learning as language use: A cross-linguistic model of child language development. *Psychological Review*, *126*, 1–51. https://doi.org/10.1037/rev0000126

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97. https://doi.org/10.1037/h0043158

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*, 39–41. https://doi.org/10.1145/219717.219748

Naiman, N. (1974). The use of elicited imitation in second language acquisition research. *Working Papers on Bilingualism*, *2*, 1–37.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners* (Unpublished doctoral dissertation). University of Hawai'i at Manoa, Hawai'i, United States.

Papasalouros, A., Kanaris, K., & Kotis, K. (2008). Automatic generation of multiple choice questions from domain ontologies. *Proceedings of IADIS International Conference e-Learning* (pp. 427–434). Amsterdam, Netherlands: IADIS.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318). Philadelphia, PA: Association for Computational Linguistics.

Rebuschat, P., & Williams, J. (2012). *Statistical learning and language acquisition*. Berlin, Germany: De Gruyter Mouton.

Scott, M. L., Stansfield, C. W., & Kenyon, D. M. (1996). Examining validity in a performance test: The listening summary translation exam (LSTE)-Spanish version. *Language Testing*, *13*, 83–109. https://doi.org/10.1177/026553229601300106

Slobin, D. I., & Welsh, C. A. (1967). Elicited imitation as a research tool in developmental psycholinguistics. In C. Ferguson & D. I. Slobin (Eds.), *Studies of child language development* (pp. 485–497). New York, NY: Holt, Rinehart & Winston.

Stanford. (2001). Stanford WebBase project. Retrieved from http://dbpubs.stanford.edu:8091/˜testbed/doc2/WebBase/

Studdert-Kennedy, M. (1986). Some developments in research on language behavior. In N. J. Smelser & D. R. Gerstein (Eds.), *Behavioral and social science. Fifty years of discovery: In commemoration of the fiftieth anniversary of the "Ogburn Report: Recent Social Trends in the United States"* (pp. 208–248). Washington, DC: National Academy Press.

Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, *38*, 1229–1261. https://doi.org/10.1017/S014271641700011X

Suzuki, Y., & DeKeyser, R. (2015). Comparing elicited imitation and word monitoring as measures of implicit knowledge. *Language Learning*, *65*, 860–895. https://doi.org/10.1111/lang.12138

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, *30*, 415–433. https://doi.org/10.1177/107769905303000401

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–336.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 279–298). Amsterdam, Netherlands: John Benjamins.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research: "Clozing" the gap. *Studies in Second Language Acquisition*, *33*, 339–372. https://doi.org/10.1017/S0272263111000015

Van den Noort, M. W., Bosch, P., & Hugdahl, K. (2006). Foreign language proficiency and working memory capacity. *European Psychologist*, *11*, 289–296. https://doi.org/10.1027/1016-9040.11.4.289

Wu, S., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annuals*, *46*, 680–704. https://doi.org/10.1111/flan.12063

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, *33*, 497–528. https://doi.org/10.1177/0265532215594643

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Spanish Recall Task Materials.
Appendix S2. English Recall Task Materials.
Appendix S3. Translation Task Materials.
Appendix S4. Demographic and Language Questionnaire.

Appendix: Accessible Summary (also publicly available at https://oasis-database.org)

Your Recall of Spoken Utterances From a Second Language Tells Us About Your Proficiency in That Language

What This Research Was About and Why It Is Important

Determining how proficient someone is in a second language (L2) is difficult and often time-consuming. Moreover, many of these tests may not provide good indications of how the learner will fare in real-world conversations. This study explored whether recalling a few spoken utterances taken from television shows can tell us something about how proficient people are in an L2. The results suggest that people's ability to recall such utterances provides a good indicator of how well they can understand spoken sentences in the L2.

What the Researchers Did

- Ninety Spanish-learning participants, with different degrees of proficiency, participated in this study.
- The participants listened to five short Spanish audio clips (5–7 seconds long) and then recalled them immediately thereafter from memory. They also recalled five short English auditory clips of similar length.
- The participants completed a shortened version of the *National Spanish Exam* (NSE), a standard multiple-choice test used in many classrooms in the United States.
- Finally, the participants were asked to translate 18 short Spanish audio clips into English to measure language proficiency. This was meant to simulate the requirements of understanding L2 in real-time.
- All audio clips were excerpted from television shows on Netflix. Additionally, scoring of recall performance was automated via a computer program.

What the Researchers Found

- Participants' proficiency, as measured by their translation skill, was strongly associated with their ability to recall utterances in Spanish, their L2.
- English recall and performance on the NSE were not as strongly associated with translation skill.
- The computer-based scoring of performance worked almost as well as more traditional human evaluations.

Things to Consider

- The results suggest that utterance recall in an L2 may provide a useful measure of proficiency in that language, while also being easy to administer and score.
- Recall taps into people's ability to process L2 input in real-time, as needed in a conversation, suggesting that this task may be useful for assessing proficiency as related to everyday interaction.
- It is important to note that utterance recall does not provide a measure of all aspects of L2 learning skill, such as reading competence and the ability to produce utterances by oral or written means.
- The accuracy of the recall measure may be further improved by carefully varying the vocabulary and grammatical complexity of the audio clips.

Materials and data: Materials and data are publicly available at https://osf.io/guem7/.
How to cite this summary: Culbertson, G., & Christiansen, M. H. (2020). Your recall of spoken utterances from a second language tells us about your proficiency in that language. *OASIS Summary* of Culbertson et al. in *Language Learning*. https://oasis-database.org