ISABELLE DARCY, BRIAN ROCCA, ZOIE HANCOCK, SEUNG SUK LEE *

Do Individual Differences in Stress Perception and in Selective Attention Relate to Improvements in Spontaneous Speech?

Abstract. In the present paper, we examine the question of whether individual differences in lexical stress perception and in selective attention relate to improvements in spontaneous speech. We measured pronunciation improvements in 25 L2 learners of English enrolled in a 7-week oral communication course. Improvements in overall comprehensibility and in word stress realization were measured through ratings of learners' speech recordings. Their ability to perceive and encode word stress placement in English was measured with an auditory phonological judgment task; selective attention was evaluated using the Attention Network Test. The results reveal that individual differences in word stress encoding are moderately related to the speech measures, but selective attention is not.

1. Introduction

A growing number of studies examine how individual differences in perception or cognitive abilities impact second language (L2) phonological processing. Studies have shown that more accurate perception seems to correlate with more accurate production and that higher executive functioning (e.g. attention, inhibition, working

* **Addresses for correspondence**: Prof. Dr. **Isabelle DARCY**, Indiana University, Dept. of Second Language Studies, Morrison Hall 231, 1165 E. 3rd St., BLOOMINGTON, IN 47405, USA
E-mail: idarcy@indiana.edu
Research areas: pronunciation instruction, second language phonology, the bilingual mental lexicon.

Brian ROCCA, Ph.D. student, Indiana University, Dept. of Second Language Studies, Morrison Hall 231, 1165 E. 3rd St., BLOOMINGTON, IN 47405, USA
E-mail: brocca@iu.edu
Research areas: pronunciation instruction, second language phonology.

Zoie HANCOCK, MA student, University of Hawai'i at Mānoa, Dept. of Second Language Studies, Moore Hall 570, 1890 East-West Road, HONOLULU, HAWAI'I 96822, USA
E-mail: zhancock@hawaii.edu
Research areas: pronunciation instruction, second language phonology.

Seung Suk LEE, Ph.D. student, University of Massachusetts Amherst, Dept. of Linguistics, Integrative Learning Center, 650 N. Pleasant St., AMHERST, MA 01003, USA
E-mail: seungsuklee@umass.edu
Research areas: pronunciation instruction, phonological learning, phonetics, Korean linguistics.

DOI 10.2357/FLuL-2020-0017 FLuL 49 (2020) • Heft 2

memory) seems to correlate with higher fluency and more accurate pronunciation or phonological processing.

However, few studies have examined how such individual differences might relate to learners' improvements as a result of pronunciation instruction. In the domain of L2 pronunciation, the ability to focus one's attention combined with more accurate phonological processing could provide learners with larger benefits from instruction. This could be potentially important for understanding which instructional strategies are effective for particular learners depending on their specific abilities (as suggested, for example, by the Aptitude-Treatment Interaction framework, SNOW 1989).

1.1 More accurate perception of a phonological dimension can benefit production

Many L2 speech learning studies have shown an interest in understanding whether the perception and production modalities are related, and if so, how. This relationship is a central tenet of one of the major theoretical models of the field, the Speech Learning Model (cf. FLEGE 1988, 1995), which assumes that the accuracy with which non-native sounds are produced is in part determined by perceptual accuracy.

Across studies that examined whether accurate perception correlates with accurate production of a phonological dimension, however, the picture is mixed. Some find that performance in perception and production of a contrast can be dissociated, showing no correlation in performance between perception and production modalities (cf. e.g. DARCY/KRÜGER 2012; KARTUSHINA/FRAUENFELDER 2014; PEPERKAMP/BOUCHON 2011), while others find evidence that there is such a relationship (cf. e.g. FLEGE 1993; FLEGE/BOHN/JANG 1997; KLUGE et al. 2007; SEBASTIAN-GALLES/BAUS 2005). Interestingly, effects showing a relationship at the level of experimental groups tend to disappear when individuals are considered (cf. also KARTUSHINA/FRAUENFELDER, 2014). A study by BRADLOW and colleagues (1997) suggests that perception training of a non-native contrast can improve the production of that contrast, as shown by the fact that production improvements were observed in a group of learners who received perceptual training, but not in the control group (no perceptual training).

However, after analyzing the individual improvement data to test the hypothesis that the subjects who show the most perceptual learning will also show the most production improvement, no correlation between perception and production was found. This led the authors to conclude that "it is not the case that improvement in perception and production proceeded in parallel within individual subjects" (BRADLOW et al. 1997: 2307).

Generally however, it seems to be the case that perceptual training can help L2 learners to improve both their perception and production of segmentals and suprasegmentals (cf. BRADLOW/AKAHANE-YAMADA/PISONI/TOHKURA 1999; LEE/LYSTER 2017; WANG/JONGMAN/SERENO 2003). A recent meta-analysis examined whether perception training can lead to improvements in production accuracy. The findings

converge to suggest that "perception training affords medium-sized gains on perception, and the experimental groups experience a small but trustworthy improvement in their production after perception training" (SAKAI/MOORMAN 2018: 204). However, this relationship may be weaker at the individual level.

1.2 Higher executive functioning can benefit production, too

In a similar vein to the relationship between perception and production, extensive research has uncovered a number of factors underlying differences in individual attainment among (mostly instructed) L2 learners. Of course, certain conditions of learning (such as the learners' first language [L1], or their age and length of learning) are known to affect pronunciation accuracy and the degree to which pronunciation is comprehensible. However, individual differences often remain after these variables are controlled (cf. BRADLOW et al. 1997; GOLESTANI/ZATORRE 2009).

In the realm of cognitive abilities and executive control, three main components of "executive functions" have been found to relate to language learning outcomes: working memory (cf. MIYAKE/FRIEDMAN 1998), attention (cf. SEGALOWITZ/FRENKIEL-FISHMAN 2005), and inhibitory control (cf. MERCIER/PIVNEVA/TITONE 2014). These components have been associated with higher proficiency or more efficient processing, and overall, higher abilities in executive control have also been found to relate to more accurate pronunciation and phonological processing (cf. e.g. ALIAGA-GARCIA/MORA/CERVIÑO-POVEDANO 2011; DARCY/MORA/DAIDONE 2016; DARCY/PARK/YANG 2015; HU et al. 2013; LEV-ARI/PEPERKAMP 2013). However, this line of research has mostly evaluated isolated dimensions of phonological systems or their specific acoustic-phonetic properties (e.g. vowel categorization, voice onset time) and not global characteristics of L2 speech, such as comprehensibility or accentedness, although limited evidence suggests that more effective attention control may be related to increased L2 fluency (cf. GARCIA-AMAYA/DARCY 2013) and comprehensibility (cf. MORA/DARCY 2017).

One area in particular – selective attention – is relevant for the present study: various perspectives suggest a relationship between pronunciation accuracy and directing attention towards phonological dimensions (cf. also SCHMIDT 1990). For example, instructional methods in which learners' attention is drawn to phonological dimensions, such as explicit pronunciation instruction (cf. DERWING/MUNRO/WIEBE 1998; GORDON/DARCY/EWERT 2013; PENNINGTON/ELLIS 2000), or corrective feedback (cf. HARDISON 2004; SAITO 2011), lead to measurable improvement. Thus, it is possible that learners who can 1) focus attention on speech dimensions in the input or in their own productions (noticing), and 2) inhibit irrelevant information (such as L1-interference) at the same time, are better equipped to benefit more from explicit instruction techniques (cf. TROFIMOVICH/GATBONTON 2006). In fact, a recent dissertation (cf. GÖKGÖZ-KURT 2016) found that the more efficient the attention control, as measured by one verbal and one non-verbal task, the more gains were made between the pre- and post-test on a connected speech perception task. However, this study did not

examine global pronunciation patterns such as comprehensibility, nor did it examine phonological perception separately.

In the current study, we examine whether individual differences in lexical stress perception and in selective attention relate to comprehensibility improvements observed in students after seven weeks of instruction in an oral communication course.

2. The present study

2.1 Method

We measured pronunciation improvements in 25 L2 learners of English enrolled in a 7-week oral communication course. Improvements in overall comprehensibility and in word stress realization were obtained through ratings of spontaneous speech recordings made at the beginning and end of the course, six weeks apart (see 2.1.2). We also measured their ability to perceive and encode word stress placement in English with an auditory lexical decision task (e.g. *girAFFE* (real word, with second syllable stress) vs. **GIRaffe* (non-word) see 2.1.5); additionally, we measured their selective attention using the Attention Network Test (ANT), specifically the flankers component (see 2.1.4; cf. FAN et al. 2002; WEAVER et al. 2009; WEAVER/BÉDARD/MCAULIFFE 2013) and the derived conflict effect (see 2.2.2 for more details). Language learning history and demographic information about participants were obtained through a background questionnaire.

2.1.1 Participants

A group of 25 L2 learners of English and 5 native speakers of English participated in the study. The learners were from six different intact classes in an Intensive English program at a large Midwestern university in the U.S., and had various L1s. All students were enrolled in an oral communication course, not a pronunciation-focused course, lasting 7 weeks. Of the six classes, four received explicit pronunciation feedback and some form of pronunciation training (the role of this training and feedback on speech samples is discussed in a separate manuscript; cf. DARCY/ROCCA/HANCOCK/LEE *in preparation*); however, the sample size does not allow us to examine here the effect of instruction on individual differences. In addition, while instruction differences may ultimately affect the outcome measures we examine, they are unlikely to impact the stress perception or the selective attention measures since these tests were performed in the first or second week. Actual enrollment numbers were slightly higher, but only learners who participated in both the pre-test and the post-test were included in the analyses (see 2.1.2 and Table 1). The native speakers were enrolled as graduate students at the same institution at the time of the study. They all reported being exposed only to American English (from various areas) or Canadian English (British Columbia) from infancy, and have remained dominant in that language since. They all knew additional languages learned after the age of 10.

 DOI 10.2357/FLuL-2020-0017

	Learners	Native Speakers
N	25*	5
Mean age, SD	26.5 (7.2)	32 (6.2)
Mean length of learning English (years), SD	11.9 (5.8)	N/A
Gender (m, f)	(15, 21)	(2, 3)
Mean age of first exposure to English, SD	14.7 (8.9)	N/A
Mean self-estimated proficiency (added across all four skills), SD	6.2 (1.5)	N/A
Mean number of L2s spoken, SD	1.3 (0.7)	1.4 (0.5)

Note. The "mean number of L2s spoken" is the mean number of L2s (including English) spoken by each participant. *25 learners provided speech samples at two times, but 35 learners were enrolled across the six classes. Some fields were not applicable (N/A) for native speakers.

Table 1: Group sample size in the study and demographic information

2.1.2 Speech samples

Speech samples were recorded in week 1 (pretest, T1) and in week 6 (post-test, T2) from learners. We used 3 tasks to measure changes from pre- to post-test (see Table 2) adapted from HASLAM (2017). To obtain recordings that were as natural and as high-quality as possible, and to encourage students to forget about the microphones, we equipped each student with a Polsen PL-2WC directional Cardioid Lavalier microphone and a Zoom ZH1 recorder on the days of the recordings. Students were recorded in their classroom, separated from each other for individual tasks (1 and 3), and divided into groups for the group task (2). The order of tasks was variable between the pre- and the post-test. Each student was provided with a handout containing instructions and the materials for the speaking tasks.

1: Controlled reading task	Reading the "Jones Family" passage (recorded individually)
2: Group discussion task	Problem solving in an information gap task – recorded individually but performed in small groups (open a restaurant, find an apartment)
3: "Transfer" speaking task	Speaking spontaneously on a given topic such as travel, holidays, food, etc. – recorded individually

Table 2: Overview of tasks used to obtain the speech recordings

The five native speakers of English were also recorded while completing the same tasks, only once, using the very same equipment and handouts as the L2 learners. Their recordings were handled in the same way and mixed with the learners' in the rating tasks.

Participants productions were orthographically transcribed. Twelve comparable sentences (based on length and complexity) were then selected among the transcripts

by an assistant unaware of the purpose of the sentences or of the experiment (and who did not hear the samples) for all 25 students (6 for T1 and 6 for T2 per student). Six sentences for each of five native speakers were also extracted. Two sentences for a student were excluded because noise levels were too high in the entire recording for that student, resulting in a total of 328 audio samples.

2.1.3 Rating of speech samples

All samples obtained for a given task (reading, group, speaking) were randomly ordered and presented to 5 trained raters who rated the comprehensibility of each item. Comprehensibility was defined as ease of understanding on a 1-9 scale (1: very easy to understand; 9: impossible to understand). Raters also marked specific errors pertaining to word stress and vowel reduction on the transcript of each item. Sentences were grouped by task to facilitate the rating task. To keep the length of a given rating session manageable and to allow for breaks, blocks of about 55 sentences were created, two per task, and resulted in six blocks of sentences for the word stress rating, and another six blocks of sentences (in a different random order) for the vowel reduction rating. The sentences for reading were presented first, followed by the group and finally the speaking task.

During each rating block, raters heard a sentence and first rated its comprehensibility. Then, they marked word stress errors on the transcript of the sentences. After all samples had been rated once, a new set of rating blocks with a new random order was presented for vowel reduction errors. Again, upon hearing a sample, raters rated its comprehensibility and then marked vowel errors. They were allowed to listen to each sample up to 8 times. All raters first rated word stress, and then vowel reduction. Comprehensibility was rated in both rating tasks. All raters were trained beforehand on unrelated sentences to determine, for example, a word-stress error in a multisyllabic word vs. a vowel reduction error in function words.

Each rating task produced 6 datasets per rater (2 for each task). Due to a computer error, one dataset for one rater was not saved, resulting in 59 datasets instead of 60. Ratings were administered with the Praat software, version 6.0.52 (cf. BOERSMA/WEENINK 2019). Overall, completing all ratings took 11-12 hours per rater, and ratings were spread over 6 or more sessions of 2 hours at most. Frequent breaks were encouraged. Raters performed the rating task individually in a quiet research lab on a university campus and were paid for their time. Given that the same order was presented to all raters, the possibility of sequence effects on the ratings exists. However, the lack of any clear relationship between the order number and the proportion of errors raters noted ($r = .007$ for word stress order; $r = -.119$ for vowel reduction order) suggests that the possibility is low (if not zero).

Comprehensibility ratings were averaged for each student/time/task. For word stress, scores were extracted by counting the number of errors marked by raters for each student at each time in each task, and expressed as a ratio of the total number of multisyllabic words and compound nouns in a given person's samples. This number

was the same in the reading samples but differed in both other speaking sample types that were more spontaneous. A higher ratio corresponds to more errors. For vowel reduction, the number of syllables in each individual sample was automatically extracted. Since essentially any word can produce an error (including function words), the scores were calculated as a proportion (in %) from the total tallied errors by all raters divided by the total number of syllables in the transcript of each sample (multiplied by the number of raters).

While the rating was long, all raters expressed confidence in their rating. The inter-rater reliability scores for each rating task (comprehensibility, word stress, vowel reduction) were high. The intraclass correlation coefficient (ICC), which reflects both the degree of correlation and agreement between measurements was used for the analysis of interrater reliability among the raters for each group of ratings (comprehensibility; vowel reduction errors; word stress errors). ICC estimates and their 95% confident intervals were calculated using SPSS 26 based on a mean-rating (k = 5), consistency, 2-way random-effects model.[1] For the comprehensibility and vowel reduction models, the inter-rater reliability was excellent, for word stress, it was good: Comprehensibility: ICC(2, 5) = .901 [.888 - .913]; Vowel-reduction: ICC(2, 5) = .913 [.896 - .928]; Word stress: ICC(2, 5) = .896 [.877 - .913].

2.1.4 Selective attention task: flankers

To obtain a measure of selective attention, we used the short ANT task using flankers without the orienting cues (cf. WEAVER/BÉDARD/MCAULIFFE 2013). At each trial, participants were shown an image of five arrows in the middle of the screen and were asked to press either the left or right arrow key on the keyboard to match the direction of the arrow in the center, ignoring all the adjacent arrows (the flankers). There were two types of images, where the arrow sequence was congruent or incongruent. Congruent images had all the arrows pointing in the same direction (all left or all right, e.g. → → → → →). Incongruent images had the center arrow pointing opposite the flanking ones (e.g. → → ← → →). All images had 5 arrows each and the arrows were the same size, color, and location in the picture.

Most participants were tested individually in a quiet research lab on a university campus. A small number of them were tested in a quiet computer lab, at individual workstations separated by partitions. They were seated at a desktop computer, wearing high quality headphones. Written instructions were displayed on the screen and orally supplemented by the experimenter following a defined protocol. Speed was emphasized. After an 8-item trial session, with feedback indicating reaction time (RT) to encourage speeding up responses, participants started the experimental block, where no feedback was provided. On each trial, a fixation cross was first shown, followed by one of the four images. Participants then indicated their answer using the arrow keys on the keyboard. The software PsychoPy (cf. PEIRCE 2007) was used to

[1] Cf. KOO/LI 2016 for justification of these choices.

present stimuli and record responses. The RT from the beginning of the trial until the key-press was measured, as well as accuracy.

2.1.5 Stress perception

To evaluate learners' ability to perceive word stress and associate it with English word forms, we used an auditory phonological judgment task which involved deciding whether a spoken word form was the expected pronunciation for a picture presented on the screen. Specifically, participants were asked to judge whether the pronunciation they heard was the way native speakers of American English usually pronounced the word. If they thought the audio stimulus they heard corresponded to a common, expected pronunciation of this word in American English, they pressed the right arrow; if not, they pressed the left arrow key on the keyboard.

2.1.5.1 Materials and conditions

There were two conditions: the test condition (condition S), where the difference between expected and unexpected (*) was in the stress placement (e.g. respectively girAFFE vs. *GIRaffe); and the control condition (condition P), where the difference between expected and unexpected (*) was in phonemes, either vowels or consonants (e.g. board vs. *noard, for a picture of a blackboard). For each condition, 28 frequent and picturable words were chosen; 14 were modified for stress, and 14 for phonemes, resulting in 28 test pairs, and 28 control pairs, and a total of 112 unique audio stimuli.

The test stimuli consisted of 14 di- and 14 trisyllabic words; half of the words were stressed on the first syllable, the other half on the second. The stress placement was then modified to either the second (e.g. Canada [ˈkænədə] – *[kəˈnædə]), or the first syllable (e.g. Japan [dʒəˈpæn] – *[ˈdʒæpən]). It was never placed on the last syllable. Vowel reduction such as schwa was also adjusted, along with the word stress. The control stimuli contained more varied changes in phonemes: 21 had a consonant variation (e.g. eggs vs. *etts) and 7 had a vowel variation (e.g. plane vs. *plone). There were 15 mono-, 9 di- and 4 trisyllabic words. The audio files were recorded in a sound-isolated recording booth by a phonetically trained female native speaker of English. Individual stimuli were cut and normalized for amplitude.

Two pictures were chosen for each pair, so that both versions of a word would not be paired with the same picture. For instance, the expected and the unexpected pronunciation of giraffe did not appear with the very same picture, but two different pictures of a giraffe. Pictures were selected from the website Pixabay. Care was taken to select the most unambiguous and simple pictures available. To limit the influence of the pairing between stimulus and picture on the responses, two lists were created. In list 1, the expected stimulus was paired with picture 1, and the unexpected with picture 2. In list 2, this pairing was reversed: expected with picture 2, unexpected with picture 1. Participants were assigned to one or the other list. All 112 stimuli were split into two blocks, so that each member of a pair occurred in a different block. Stimuli within

each block, and the order of blocks were randomized by the experimental program PsychoPy (cf. PEIRCE 2007).

2.1.5.2 Procedure

Participants were tested immediately following the Flankers task, in the same conditions. Before the experiment, each participant was pseudo-randomly assigned to a list (1 or 2) by the experimenter. After clarifying any questions about the instructions, a short practice session with three pairs (containing both expected and unexpected versions) was provided with feedback to familiarize participants with the task. All practice pairs had a phonemic change. Feedback emphasized both accuracy and speed, and it provided the RT to encourage speedy responses. Instructions (see fig. 1) stressed that participants should focus on whether the pronunciation was expected or unexpected in typical American English realizations of this word.

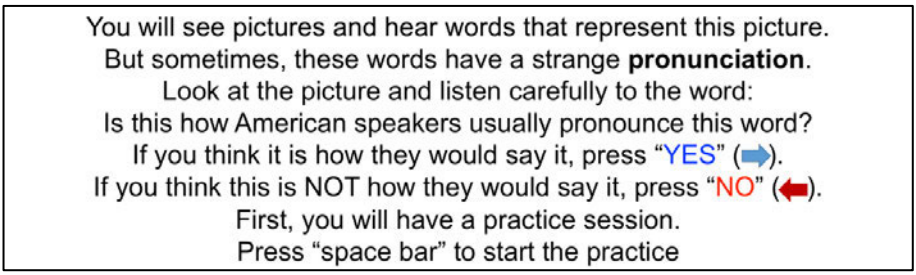

Fig. 1: Instructions screen for the phonological judgment task

After completing the practice, participants started the experimental blocks. Each trial started with a 300 ms fixation cross. The object image was first presented on the screen for 1000 ms. The audio stimulus was then played. Participants had 3000 ms to respond before the script moved to the next trial. There was a 300 ms inter-trial interval (blank screen). Participants had to indicate their answer on the keyboard using the arrow keys and no feedback was provided. There were two experimental blocks of 56 trials each. Between blocks, participants were given time for a break. The total experiment took approximately 12 minutes to complete. Accuracy and RTs (from the end of the sound file) were measured.

2.2 Results by task

2.2.1 Speech samples comprehensibility, vowel reduction and word stress

For each participant, six scores were generated from the ratings: comprehensibility at T1, at T2; vowel reduction errors at T1, T2; and word stress errors at T1, and T2. These six scores were also converted into "gain" scores, that is, the difference between

T1 and T2. For example, a participant whose vowel error score at T2 is 4, and at T1 was 6, has a gain score of -2 points, indicating that this learner has reduced their errors by 2 points. Negative gains correspond to improvement, as they indicate a reduction in errors, or a smaller comprehensibility score. Recall that a score of 1 corresponds to highest comprehensibility on our task. Table 3 shows central tendencies and the amount of variability for the three measures at T1, T2, and the gain scores.

Measure	**Mean T1 (*SD*)**	**Mean T2 (*SD*)**	**Mean gain, T2-T1 (*SD*)**
Vowel reduction errors	8.06 (2.85)	7.21 (3.59)	-0.85 (2.47)
min - max	2.30 – 13.15	1.94 – 14.23	-6.53 – 4.37
Word stress errors	19.68 (8.26)	18.33 (10.14)	-1.35 (6.58)
min - max	7.22 – 36.58	4.46 – 38.61	-2.76 – 16.67
Comprehensibility	4.62 (1.00)	4.52 (1.16)	-0.10 (0.53)
min - max	2.78 – 6.52	2.36 – 6.72	-1.07 – 1.53

Note. Comprehensibility scores are on a scale of 1 to 9, where 1 is most comprehensible. Vowel and word stress errors are proportions converted to percentages.

Table 3: Central tendencies, minimum and maximum scores for the learners at T1, T2, and for the gain scores (T2-T1), in each measure

2.2.2 Flankers

Data for one participant were excluded because this person did the task twice. The final sample is n = 35 for this experiment (learners n = 30; native speakers n = 5). Only RT for correct answers were considered. Further, all datapoints that were either under 300 ms (n = 0) or beyond 2 *SD* from the mean RT were discarded (n = 40 datapoints, or 3% of the total dataset).

The mean RT for congruent trials was 504 ms (*SD* = 91), and the mean RT for incongruent trials was 529 ms (*SD* = 81), indicating that responses were significantly slower in incongruent trials, $t(34) = -3.98$, $p < .001$, as expected. The mean RT for each congruency condition for each person was used to compute the conflict effect as the difference in RTs between congruent and incongruent trials (cf. BUGG/CRUMP 2012) (*M*_incongr. – *M*_congr. = conflict effect). A smaller conflict effect indicates a more efficient control of selective attention (*M* = 24.5; *SD* = 36.5). Values ranged from -58 to +135 ms, with a similar mean and variance in both native speakers and learners. Conflict effect scores were used in the correlation analysis below.

2.2.3 Stress perception

Because error rates can depend on a person's response strategy (saying 'yes' all the time to both expected and unexpected forms, for example), it is useful to compute a sensitivity measure that takes such biases into account (cf. MACMILLAN/CREELMAN

 DOI 10.2357/FLuL-2020-0017

2005: 8ff.). Therefore, a *d'* (dee-prime) measure of sensitivity was calculated from the responses. Answers on each trial were first coded into Hits ("yes" for expected), False Alarms ("yes" for unexpected), Misses ("no" for expected) and Correct Rejections ("no" for unexpected), in order to compute the *d'* score for each participant in each condition (stress vs. phoneme). *d'* scores are calculated by comparing the rate of *hits* with the rate of *false alarms,* converted into *z*-scores (standard deviation units). When participants cannot discriminate at all and respond randomly, hits = false alarms and *d'* = 0. The higher the hit rate compared to false alarms, the higher the *d'* and the higher the sensitivity. A *d'* value below 0.75 can be interpreted as a lack of discrimination sensitivity, whereas values at or above 3.0 show a very strong discrimination sensitivity. The data on the control (phoneme) condition was then screened for outlier performance, which might indicate that a participant did not know a reliable number of words or did not understand the task. For learners, the mean accuracy on this condition was 86% (*SD* = 12). Three outliers were detected, whose performance was below 2 *SD* from the mean on this condition in the learner group, and were excluded from further analyses. No outlier was detected among the native speakers (*M* = 99%, *SD* = 1.6). The resulting final sample is 29 learners, and 5 native speakers. Figure 2 illustrates the mean *d'* for each condition in each group (learners, native speakers). It becomes clear that the learner group is much more variable in performance on stress condition, with *d'* scores ranging from > 3 (indicating very high sensitivity) to < 1 (indicating limited sensitivity). For the other condition, no *d'* was below 2. Similarly for the native speakers, sensitivity is very high regardless of condition.

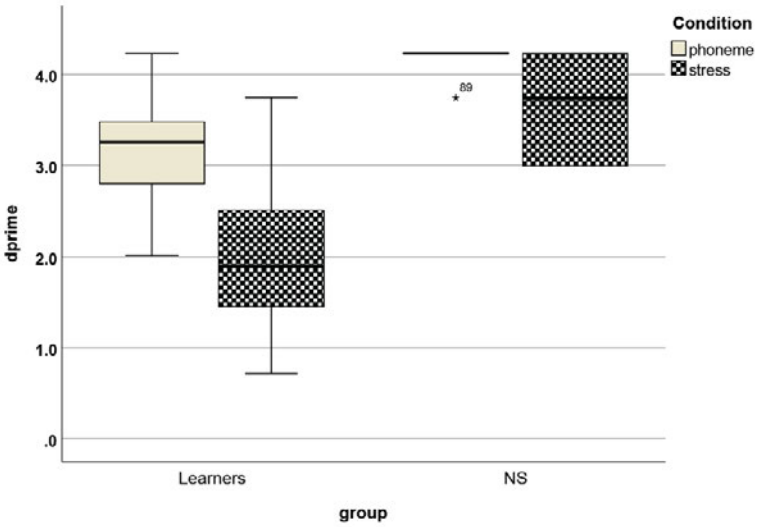

Fig. 2: Boxplot of *d'* score averages in each condition (phoneme, stress) for learners and native speakers

A correlation was conducted to estimate how strongly the learners' performance in one condition was linked to the other condition. No clear relationship emerged ($r(29) = .209$, $p = .277$). Therefore, the *d'* scores on the stress condition, which reflect individual performance in word stress lexical encoding, will be used in the correlation analysis.

2.3 Correlations

The main variables of interest in each task are as follows. For comprehensibility (COMP), word-stress (WS) and vowel reduction (VR) errors, we examine scores at T1 and T2, and the difference between T1 and T2 (GAIN). For the attention task, we use the conflict effect, indexing selective attention. For the phonological judgment task, we use the *d'* score for the stress condition (S-dprime), indexing the extent to which a learner perceives and encodes lexical stress. Except where otherwise indicated, two-tailed Pearson correlations were conducted with only the participants who were not outliers in any task. Native speakers are not included. The final sample is n = 22.

2.3.1 Speech measures correlations

First, the relationship of the speech measures at T1 and T2 to one another was examined. All were highly correlated with each other, which suggests that performance at T1 related to each person's performance at T2. In addition, both word stress and vowel reduction errors correlate with comprehensibility, underscoring their role in how easily a listener understands their speech (for WS and COMP at T2, $r = .755$, $p = .001$; for VR and COMP at T2, $r = .715$, $p = .001$; similar values were obtained for T1 between these variables).

2.3.2 Speech measures with cognitive and perception variables

In the current study, our main research question is whether individual differences in lexical stress encoding and in selective attention relate to comprehensibility as well as word stress or vowel reduction errors, or to improvements in these speech measures after an oral communication course. To answer these questions, several correlation analyses were performed in SPSS 26, between the speech measures, the gain scores, and the individual differences measures. If they are related, we expect that a higher *d'* in lexical stress encoding and a lower conflict effect are associated with both lower error rates and a lower comprehensibility score (which indicates higher comprehensibility on our Likert scale), as well as more negative gain scores (which indicate improvement). That is, we expect a negative correlation between *d'* and the speech measures. Conversely, a positive correlation is expected between the conflict score (a smaller conflict effect indicates a more efficient control of selective attention), and error rates/comprehensibility score (where smaller values indicate more target-like

 DOI 10.2357/FLuL-2020-0017

performance). Given the expected direction of the effect, the correlations were one-tailed. The results are presented in Table 4.

Measure		Conflict effect	S-dprime
T1 WS	Pearson Correlation	-.301	**-.540****
	Sig. (1-tailed)	*.087*	*.005*
T1 VR	Pearson Correlation	-.311	**-.678****
	Sig. (1-tailed)	*.079*	*.000*
T1 COMP	Pearson Correlation	-.229	-.304
	Sig. (1-tailed)	*.153*	*.084*
T2 WS	Pearson Correlation	-.148	**-.502****
	Sig. (1-tailed)	*.256*	*.009*
T2 VR	Pearson Correlation	-.216	-.309
	Sig. (1-tailed)	*.167*	*.081*
T2 COMP	Pearson Correlation	-.235	-.267
	Sig. (1-tailed)	*.146*	*.115*
GAIN WS	Pearson Correlation	.202	-.019
	Sig. (1-tailed)	*.184*	*.467*
GAIN VR	Pearson Correlation	.048	.323
	Sig. (1-tailed)	*.417*	*.071*
GAIN COMP	Pearson Correlation	-.093	-.027
	Sig. (1-tailed)	*.340*	*.453*
N		22	22

Note. WS = word stress; VR = vowel reduction; COMP = comprehensibility; S-dprime = *d'* on stress condition; ** Correlation is significant at the 0.01 level (1-tailed).

Table 4: Correlations between speech measures and individual differences measures of attention (Conflict effect) and word stress perception (S-dprime)

The results indicate that the conflict effect consistently shows no significant relationship in the predicted direction with any of the speech measures. However, the perception score (S-dprime) relates moderately and significantly, in the expected direction, with T1 WS and VR errors (r = -.540; r = -.678, both $p < .01$), as well as with T2 WS errors (r = -.502; p = .009; r^2 = .252). A similar relationship, albeit weaker, and marginally significant (r = -.309; p = .081; r^2 = .095), also in the expected direction, is found with VR errors at T2 and S-dprime. Interestingly, the relationship between stress perception and T1 speech measures seems stronger than at T2. This might be due to the effect of instruction received by some learners in our sample. We address this possibility in the discussion. There were no correlations between gain scores and either of the individual differences variables.

Finally, correlations examined whether the T1 scores might predict the gain size. But they did not. One relationship emerged: The gain scores in word stress errors were positively related with gain scores in comprehensibility, indicating that reducing word stress errors might contribute to increased comprehensibility (Figure 3). However, since correlations are not causal, such a conclusion cannot be drawn definitively.

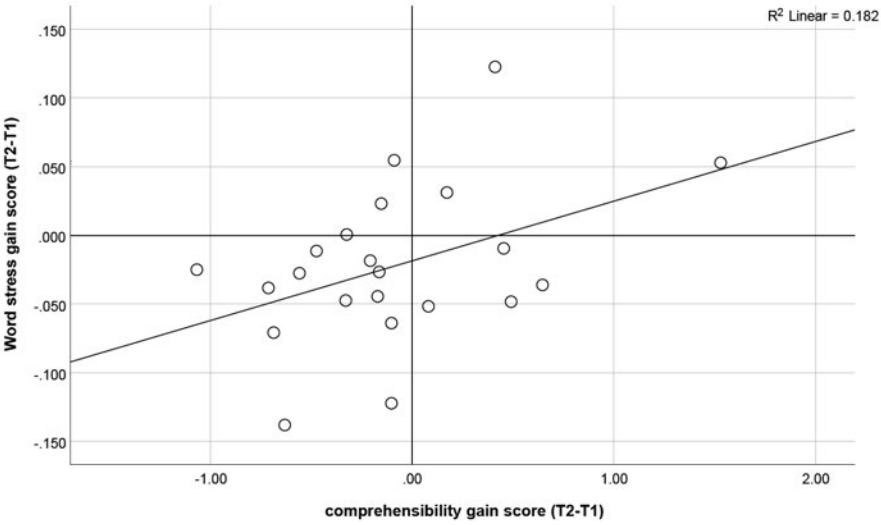

Fig. 3: Scatterplot of the gain in word stress errors (T2 – T1), and the comprehensibility gains (T2 – T1) – For both, a negative score indicates a more target-like performance, since a lower score at T2 is desirable.

3. Discussion

In the current study, we set out to explore whether individual differences in lexical stress encoding and in selective attention relate to comprehensibility improvements as well as in word stress or vowel reduction errors after an oral communication course. Our results indicate that in our sample, selective attention consistently did not relate to any speech measure at T1 or T2, or to the size of the gain between T1 and T2. The absence of relationship in our relatively small sample may not be surprising. While it could emerge in a study with a much larger sample, its strength may still be moderate.

A clearer relationship emerged between the lexical stress encoding ability (indexed by the *d'* score on the stress condition) and several speech measures at T1, and also, albeit weaker, at T2. The strongest relationship was found between lexical stress *d'* and the vowel reduction errors at T1. Similar, but weaker relationships were also observed at T2 with word stress and vowel reduction errors, and with comprehensi-

bility – where it did not reach significance however. Again, significance levels are likely a matter of power, but it is striking that lexical stress encoding is more clearly related to the speech measures at T1 than at T2. The reasons for this phenomenon are not immediately clear, yet we think it is possibly the case that the relationships at T2 were blurred as a result of instruction during the course. Recall that four classes received instruction and/or feedback on the pronunciation of word stress and vowel reduction among others. It is therefore possible that the initial relationship we observed at T1 became less relevant or defining of a given student's performance at T2 as a result of instruction. Our sample size does not allow us to explore this development. Future research is definitely needed to understand how instruction interacts with the relationship between stress perception and speech measures.

Another avenue to explore the factors that modulate the relationship between lexical stress encoding and comprehensibility is the L1 background of the participants. In our sample, learners had different L1 backgrounds (Arabic, Chinese, and Korean were the most common ones, and there was one Spanish student). It is possible that, even though all students had difficulties with producing lexical stress and vowel reduction on the appropriate syllables in English words, some students might have fewer issues because of their L1, such as Spanish students. More experimental evidence taking this aspect into account is needed, and would contribute to generating a comprehensive overview of the relative importance of pronunciation areas for increasing comprehensibility based on the learners' language backgrounds.

Clearly though, even when instructional effects are not yet visible (at T1), individual differences in stress perception are correlated to speech outcomes; they remain so (even if less) at T2. While the consistency of this relationship should be verified with a larger sample, these findings, taken together, paint a coherent picture. It seems to be the case that word stress and vowel reduction errors both contribute to comprehensibility scores. In turn, these are associated to performance in a lexical phonological judgment task comparing word-stress differences. This relationship thus indicates that someone's ability to produce word stress (and by extent, vowel reduction in unstressed positions) in English words and sentences is related to their ability to memorize the stress pattern of English words they learned. The presence of such a relationship highlights the usefulness for teachers to address the stress pattern of words explicitly when teaching new words.

In terms of the gain between T1 and T2, no clear picture emerged from our investigation, and no clear link to either individual difference variable was found. More research with more participants in classrooms is sorely needed here. In our case, it should again be noted that instructional differences also affected how much learners improved over time, and possibly more so than individual differences at the outset. While it would be very interesting to compare different types of instruction and the role of individual differences as a function of instructional method, the current sample was clearly too small to address this question. It is also possible that the fairly short duration of the course was too short to generate large differences in gain scores, which might affect the visibility of relationships between the speech measures.

In conclusion, our results indicate that improvements in spontaneous speech are moderately related to the accuracy of lexical stress encoding.

Acknowledgments

The authors wish to thank the teachers Lynn McKenzie and Sofiya Asher, our raters Danielle Daidone, Ryan Lidster, John H.G. Scott, Sadi Phillips, Tory Robinson. We are grateful to the Intensive English Program at Indiana University for their support of this work. We further thank Claudia Serrano Romero for her help with testing participants and constructing stimuli for the perception tasks, Houston McClure for his help in selecting the sentences, John Rothgerber for his expert advice on instruction implementation, and all the students who took part.

References

ALIAGA-GARCIA, Cristina / MORA, Joan C. / CERVIÑO-POVEDANO, Eva (2011): "L2 speech learning in adulthood and phonological short-term memory". In: *Poznań Studies in Contemporary Linguistics* 47, 1–14.

BOERSMA, Paul / WEENINK, David (2019): Praat: doing phonetics by computer [Computer program]. Version 6.0.52. http://www.praat.org/ (02/05/2019).

BRADLOW, Ann R. / AKAHANE-YAMADA, Reiko / PISONI, David B. / TOHKURA, Yoh'ichi (1999): "Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production". In: *Perception & Psychophysics* 61, 977–985.

BRADLOW, Ann R. / PISONI, David B. / AKAHANE-YAMADA, Reiko / TOHKURA, Yoh'ichi (1997): "Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production". In: *The Journal of the Acoustical Society of America* 101, 2299–2310.

BUGG, Julie M. / CRUMP, Matthew J. (2012): "In support of a distinction between voluntary and stimulus-driven control: A review of the literature on proportion congruent effects". In: *Frontiers in Psychology* 3, 367. https://doi.org/10.3389/fpsyg.2012.00367.

DARCY, Isabelle / KRÜGER, Franziska (2012): "Vowel perception and production in Turkish children acquiring L2 German". In: *Journal of Phonetics* 40, 568–581.

DARCY, Isabelle / MORA, Joan C. / DAIDONE, Danielle (2016): "The role of inhibitory control in second language phonological processing". In: *Language Learning* 66, 741–773.

DARCY, Isabelle / PARK, Hanyong / YANG, Chung-Lin (2015): "Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing". In: *Learning and Individual Differences* 40, 63–72.

DERWING, Tracey M. / MUNRO, Murray / WIEBE, Grace (1998): "Evidence in favor of a broad framework for pronunciation instruction". In: *Language Learning* 48, 393–410.

FAN, Jin / MCCANDLISS, Bruce D. / SOMMER, Tobias / RAZ, Amir / POSNER, Michael I. (2002): "Testing the efficiency and independence of attentional networks". In: *Journal of Cognitive Neuroscience* 14, 340–347. https://doi.org/10.1162/089892902317361886 (08/06/2020).

FLEGE, James E. (1988): "Factors affecting degree of perceived foreign accent in English sentences". In: *The Journal of the Acoustical Society of America* 84, 70–79.

FLEGE, James E. (1993): "Production and perception of a novel, second-language phonetic contrast". In: *The Journal of the Acoustical Society of America* 93, 1589–1608.

FLEGE, James E. (1995): "Second language speech learning: Theory, findings, and problems". In: *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* 92, 233–277.

FLEGE, James E. / BOHN, Ocke-Schwen / JANG, Sunyoung (1997): "Effects of experience on non-native speakers' production and perception of English vowels". In: *Journal of phonetics* 25, 437–470.

GARCIA-AMAYA, Lorenzo / DARCY, Isabelle (2013, March): "Attention control in study abroad context: longitudinal data from L2 learners of Spanish". Paper presented to AAAL (Conference of the American Association for Applied Linguistics), 16–19/03/2013, Dallas, TX.

GÖKGÖZ-KURT, Burcu (2016): *Attention Control and the Effects of Online Training in Improving Connected Speech Perception by English as a Second Language Learners* (Doctoral dissertation, University of South Carolina). https://scholarcommons.sc.edu/etd/3560/ (10/04/2020).

GOLESTANI, Narly / ZATORRE, Robert J. (2009): "Individual differences in the acquisition of second language phonology". In: *Brain and Language* 109, 55–67.

GORDON, Joshua / DARCY, Isabelle / EWERT, Doreen (2013): "Pronunciation teaching and learning: Effects of explicit phonetic instruction in the L2 classroom". In: LEVIS, John / LEVELLE, Kimberly (eds.): *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*. Aug. 2012. Ames, IA: Iowa State University, 194–206. https://apling.engl.iastate.edu/wp-content/uploads/sites/221/2015/05/PSLLT_4th_Proceedings_2012.pdf (13/04/2020).

HARDISON, Debra M. (2004): "Generalization of computer assisted prosody training: Quantitative and qualitative findings". In: *Language Learning & Technology* 8, 34–52.

HASLAM, Mara (2017): "A new comprehensive assessment tool for English pronunciation (Teaching Tip)". Paper presented at the 9th Pronunciation in Second Language Learning and Teaching conference, 01–02/09/2017, University of Utah, Salt Lake City.

HU, Xiaochen / ACKERMANN, Hermann / MARTIN, Jason A. / ERB, Michael / WINKLER, Susanne / REITERER, Susanne M. (2013): "Language aptitude for pronunciation in advanced second language (L2) Learners: Behavioural predictors and neural substrates". In: *Brain and Language* 127, 366–376. https://doi.org/10.1016/j.bandl.2012.11.006 (08/06/2020).

KARTUSHINA, Natalia / FRAUENFELDER, Ulrich H. (2014): "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation". In: *Frontiers in Psychology* 5: 1246. https://doi.org/10.3389/fpsyg.2014.01246. (08/06/2020)

KLUGE, Denise C. / RAUBER, Andréia S. / REIS, Mara S. / BION, Ricardo A. H. (2007): "The relationship between the perception and production of English nasal codas by Brazilian learners of English". In: *Proceedings of the 8th Annual Conference of the International Speech Communication Association (Interspeech 2007)*, 2297–2300. https://www.isca-speech.org/archive/archive_papers/interspeech_2007/i07_2297.pdf. (13/04/2020)

KOO, Terry K. / LI, Mae Y. (2016): "A guideline of selecting and reporting intraclass correlation coefficients for reliability research". In: *Journal of Chiropractic Medicine* 15, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012. (08/06/2020)

LEE, Andrew H., / LYSTER, Roy (2017): "Can corrective feedback on second language speech perception errors affect production accuracy?" In: *Applied Psycholinguistics* 38, 371–393. https://doi.org/10.1017/S0142716416000254. (08/06/2020)

LEV-ARI, Shiri / PEPERKAMP, Sharon (2013): "Low inhibitory skill leads to non-native perception and production in bilinguals' native language". In: *Journal of Phonetics* 41, 320–331.

MACMILLAN, Neil A. / CREELMAN, Douglas C. ([2]2005): *Detection Theory: A User's Guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

MERCIER, Julie / PIVNEVA, Irina / TITONE, Debra (2014): "Individual differences in inhibitory control relate to bilingual spoken word processing". In: *Bilingualism: Language and Cognition* 17, 89–117. https://doi.org/10.1017/S1366728913000084. (08/06/2020)

MIYAKE, Akira / FRIEDMAN, Naomi P. (1998): "Individual differences in second language proficiency: working memory as language aptitude". In: HEALY, Alice F. / BOURNE, Lyle E. (eds.): *Foreign Language Learning. Psycholinguistic Studies on Training and Retention*. Mahwah, NJ: Lawrence Erlbaum Associates, 339–364.

MORA, Joan C. / DARCY, Isabelle (2017): "The relationship between cognitive control and pronunciation in a second language". In: ISAACS, Talia / TROFIMOVICH, Pavel (eds.): *Second Language Pronunciation Assessment: Interdisciplinary Perspectives*. Bristol: Multilingual Matters, 95–120.

PEIRCE, Jonathan (2007): "PsychoPy – Psychophysics software in Python". In: *Journal of Neuroscience Methods* 162, 8–13.

PENNINGTON, Martha C. / ELLIS, Nick C. (2000): "Cantonese speakers' memory for English sentences with prosodic cues". In: *The Modern Language Journal* 84, 372–389. https://doi.org/10.1111/0026-7902.00075. (08/06/2020)

PEPERKAMP, Sharon / BOUCHON, Camillia (2011): "The relation between perception and production in L2 phonological processing". In: *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*. Rundle Mall: Causal Productions, 161–164. https://www.isca-speech.org/archive/interspeech_2011/ (10/04/2020).

SAITO, Kazuya (2011): "Examining the role of explicit phonetic instruction in native-like and comprehensible pronunciation development: an instructed SLA approach to L2 phonology". In: *Language Awareness* 20.1, 45–59. https://doi.org/10.1080/09658416.2010.540326. (08/06/2020)

SAKAI, Mari / MOORMAN, Colleen (2018): "Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research". In: *Applied Psycholinguistics* 39, 187–224. https://doi.org/10.1017/S0142716417000418. (08/06/2020)

SCHMIDT, Richard W. (1990): "The role of consciousness in second language learning". In: *Applied Linguistics* 11, 129–158. https://doi.org/10.1093/applin/11.2.129.

SEBASTIAN-GALLES, Nuria / BAUS, Cristina (2005): "On the relationship between perception and production in L2 categories". In: CUTLER, Anne (ed.): *Twenty-First Century Psycholinguistics: Four Cornerstones*. New York: Psychology Press, 279–292.

SEGALOWITZ, Norman / FRENKIEL-FISHMAN, Sarah (2005): "Attention control and ability level in a complex cognitive skill: Attention shifting and second-language proficiency". In: *Memory & Cognition* 33, 644–653.

SNOW, Richard E. (1989): "Aptitude-treatment interaction as a framework for research on individual differences in learning". In: ACKERMAN, Phillip Lawrence / STERNBERG, Robert J. / GLASER, Robert (eds.): *Learning and Individual Differences*. New York: W.H. Freeman, 13–59.

TROFIMOVICH, Pavel / GATBONTON, Elisabeth (2006): "Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction". In: *The Modern Language Journal* 90, 519–535.

WANG, Yue / JONGMAN, Allard / SERENO, Joan A. (2003): "Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training". In: *The Journal of the Acoustical Society of America* 113, 1033–1043.

WEAVER, Bruce / BÉDARD, Michel / MCAULIFFE, Jim (2013): "Evaluation of a 10-minute Version of the Attention Network Test". In: *The Clinical Neuropsychologist* 27, 1281–1299. https://doi.org/10.1080/13854046.2013.851741. (08/06/2020)

WEAVER, Bruce / BÉDARD, Michel / MCAULIFFE, Jim / PARKKARI, Marie (2009): "Using the Attention Network Test to predict driving test scores". In: *Accident Analysis & Prevention* 41, 76–83. https://doi.org/10.1016/j.aap.2008.09.006. (08/06/2020)