# MODELING CLASSROOM LANGUAGE LEARNERS' COMPREHENSIBILITY AND ACCENTEDNESS OVER TIME

## THE CASE OF L2 SPANISH

Charles L. Nagle, Iowa State University

Significant scholarship has focused on the development of L2 oral skills in naturalistic language learning. However, few studies have examined how instructed learners' pronunciation develops over time, despite the importance of the classroom context. This study addressed this gap by investigating L2 Spanish learners' comprehensibility and accentedness over a yearlong period. Twenty-six learners completed a sentence-building task on five occasions distributed throughout their second, third, and fourth semesters of college-level Spanish language instruction. Learners received 20 sets of images, combining the images in each set to form a simple sentence in Spanish. Eighteen native Spanish listeners rated learners' recordings for comprehensibility and accentedness using 9-point Likert scales, and mixed-effects models were fit to the ratings data using R. Learners were rated as quite comprehensible despite the presence of a moderate to strong foreign accent. Although both comprehensibility and accentedness improved over time, rates of change varied. Comprehensibility improved quickly but was subject to greater deceleration in rate of change over time. In contrast, accentedness improved steadily and did not exhibit the same degree of flattening as comprehensibility. These results intersect with work on naturalistic learners and suggest that pronunciation development may be characterized by phases of change.

## INTRODUCTION

To communicate successfully in the L2, learners' pronunciation needs to be readily understandable and easy to process. Yet, despite the centrality of comprehensible pronunciation to L2 communication, there is a lack of longitudinal research on how these constructs develop over time. While it appears that the greatest changes in pronunciation are achieved during immigrants' first year of intensive L2 exposure (Derwing & Munro, 2013), little is known about classroom language learners' pronunciation development even though the language classroom is an important part of most learners' experience. According to a recent Modern Language Association report (Goldberg, Looney, & Lusin, 2015), in 2013, over 1.5 million students enrolled in postsecondary language coursework in the US alone, and over 30,000 reported a primary or secondary major in languages and literatures. Understanding how classroom learners' pronunciation changes over time stands to provide an empirical basis for curricular decisions related to whether targeted training is necessary and, if so, when it should be introduced to sustain or catalyze development. Moreover, studying classroom language learners in a foreign language context will complement previous research on immigrant populations and advanced L2

speakers in immersion contexts, resulting in a more complete understanding of how L2 speakers' pronunciation changes across a range of learning scenarios and proficiency levels.

## Constructs in L2 Speech Research

L2 pronunciation scholarship has focused on global dimensions of L2 speech such as comprehensibility, fluency, and accentedness, which when taken together provide a window into how listeners evaluate L2 speakers' communicative ability. Comprehensibility represents how easy or difficult speech is to understand and can be interpreted as an index of speech intelligiblity if the latter is defined not in terms of verified transcriptions, as in previous research (e.g., Munro & Derwing, 1995), but rather as a more subjective, listener-based measure. As Saito, Trofimovich, and Isaacs (2017) point out, the operationalization of comprehensibility as intelligibility fits with real-world applications, such as language assessment scenarios where listeners are asked to provide holistic evaluations of L2 speech. In contrast to both of these constructs, accentedness is defined in reference to deviations from the pronunciation of native speech, and research has consistently demonstrated that the two are partially independent (Munro & Derwing, 1995; Trofimovich & Isaacs, 2012). Whereas accentedness judgments are predominantly tied to pronunciation-based variables, listeners seem to take into account a wider range of linguistic characteristics, including grammatical accuracy and discourse structure, when evaluating comprehensibility (O'Brien, 2014; Saito, Webb, Trofimovich, & Isaacs, 2017).

## Research on Pronunciation Development

Longitudinal research on L2 pronunciation has provided evidence of nonlinear growth while highlighting distinct developmental trajectories for comprehensibility and fluency on the one hand, and accentedness on the other (Derwing & Munro, 2013; Derwing, Munro, & Thomson, 2008; Kennedy, Foote, & Dos Santos Buss, 2015). In one of the most comprehensive studies to date, Derwing and Munro (2013) examined L1 Mandarin and L1 Slavic language speakers' pronunciation learning over a 7-year period, during which participants were living and working in Canada. The Slavic language speakers' fluency and comprehensibility improved steadily over the study, but improvements to accentedness were limited to the first two years. In contrast, no gains were observed for the L1 Mandarin group, which may have been due to the fact that the Mandarin speakers demonstrated less willingness to communicate in the L2. In another study, Kennedy, Foote, and Dos Santos Buss (2015) investigated seven L2 English speakers who were attending a university where English was the language of instruction. Although participants' comprehensibility, fluency, and accentedness improved significantly over two years, individual differences in rate of change were evident. Findings from Huensch and Tracy-Ventura's (2017) study on advanced L2 Spanish learners' fluency before, during, and after study abroad reinforce these results. Twenty-seven participants completed a picture narration task on six occasions, once before departure, three times while abroad, and twice upon their return. Recordings were transcribed and analyzed for a number of speed, breakdown, and repair fluency measures. With respect to speed, results suggest that the greatest improvements were registered at the outset of study abroad, either between the presojourn recording and the first session abroad or between the latter and subsequent data points. In contrast, a more complex pattern emerged for the breakdown measures, and no significant changes were evident for the repair variables.

To summarize, even though longitudinal research on pronunciation development has grown over the past decade, most studies have targeted advanced L2 speakers, predominantly in an

immersion context. In that context, learners' pronunciation may improve rapidly at first, such that gains are concentrated at the outset of more intensive L2 contact (Derwing & Munro, 2013). However, little is known about how classroom language learners' pronunciation changes over the first few semesters of language study, despite the fact that it is precisely during this early stage when large shifts in linguistic ability might be expected. If initial trajectories set the stage for long-term attainment, it is critical that we examine classroom learners' comprehensibility and accentedness during this formative period.

## METHOD

### Speakers

Twenty-six L1 English-speaking students were recruited from multiple sections of a second semester college-level Spanish course at a university in the US. Although all participants were late learners placed into the same course, they varied in terms of their previous Spanish experience. The mean age of onset for the group was 14.38 ($SD$ = 4.11) years, and learners reported an average of 3.35 ($SD$ = 3.17) years of Spanish instruction prior to university. Participants were tested five times over a yearlong period: two sessions during their second semester, one in February and another in April; two sessions during their third semester, one in September and another in November; one session in February during their fourth semester. The language program in which participants were enrolled followed a communicative approach to language teaching. Classes met three days per week for one hour, during which students typically completed a series of scaffolded, task-based activities in pairs. Although the program focused on developing students' speaking ability, pronunciation was not part of the curriculum. Instructors reported that they did not systematically address pronunciation, but would recast students' mispronunciations—that is, provide the correct pronunciation without explicitly drawing students' attention to it—if they interfered with communication, and students likewise indicated that they did not receive any targeted pronunciation training as part of their courses.

Five participants withdrew after the first two sessions because they decided to discontinue their study of Spanish, and data from the first session was unavailable for four learners due to technical issues. Taking into account these sources of attrition, the final data set consisted of at least four observations for 21 of the original 26 participants. Randomly missing data is not problematic for mixed-effects models, the statistical approach employed in the present study, since model estimates are based on available data points without applying listwise deletion (i.e., deleting an entire case due to one missing data point).

### Raters

Raters were 18 native Spanish speakers who were advanced L2 English speakers pursuing a graduate degree at a US university. They completed a background questionnaire eliciting information on their English ability, other foreign languages known or studied, and their familiarity with L2 Spanish. They reported a mean age of L2 English onset of 10.37 years ($SD$ = 6.64) and estimated their English proficiency to be 7.93 ($SD$ = .95; range = 6.75–9) on a 9-point scale ranging from *very poor* (1) to *extremely proficient* (9). On average, raters indicated interacting with L2 Spanish speakers about once per month (*never, n* = 6; *monthly, n* = 7; *daily, n* = 2; *more than daily, n* = 3) and assessed themselves as having moderate familiarity with the

characteristics of L2 Spanish ($M$ = 5.63, $SD$ = 2.99, range = 1–9 on a 9-point scale where 9 = *extremely familiar*). Eight raters reported linguistic training (i.e., having taken a course in linguistics) and nine had language teaching experience.

**Speaking Task**

Given that participants were novice Spanish speakers, a sentence-building task was selected for this study. Twenty sets of three images were created by combining images representing a subject, verb, and either a direct object or a location. Vocabulary was drawn from students' introductory textbook, and participants completed a computerized training module using SuperLab software prior to recording the sentences to ensure that they could easily recall the pictures associated with all vocabulary items. During the vocabulary familiarization module, participants saw an image with the word printed below it and heard the word spoken by a native Spanish speaker. After participants had reviewed the 27 picture-word pairs included in the study, they took a vocabulary quiz on which they matched images to target words by choosing the correct response from four options. Participants advanced only when they had achieved a perfect score. For the recordings, students saw one set of images per screen and were asked to form a simple sentence in Spanish based on the pictures (e.g., *Mario pinta la cocina*, 'Mario paints/is painting the kitchen'). Students were recorded individually in a sound-attenuated studio using a Shure SM10A head-mounted microphone connected to a laptop computer through an XRL to USB signal adapter.

**Rating Task**

The 20 sentences that each learner recorded at each testing session were similar in length and content given that they were elicited through sets of images as described above. Consequently, five per learner per session were randomly sampled using a randomization algorithm, and the peak intensity of each file was then normalized to 70 dB to prepare samples for rating. Thirty-five samples from seven native speakers of Spanish (i.e., five per native speaker) were also included. Audios were split into three blocks with each block containing files from all speakers at all testing sessions as well as native samples. Listeners were tested individually in a quiet space using noise-cancelling headphones. Comprehensibility was defined as ease of listening and accentedness as deviations from native speech, both on 9-point scales where 1 represented the best score and 9 the worst score. Ratings were carried out sequentially and counterbalanced across listeners to prevent ordering effects. In other words, listeners heard each file twice, once while rating each construct, with half of the raters evaluating comprehensibility first and the other half accentedness. Files were randomized within blocks such that each listener heard the clips in a unique order. The entire rating session lasted about one hour, and listeners were offered breaks in between blocks to prevent fatigue.

**RESULTS**

In this study, mixed-effects models (for an overview, Cunnings & Finlayson, 2015; Linck & Cunnings, 2015) were fit to the data using the lme4 package (Bates, Maechler, Bolker, & Walker, 2014) of R (R Core Team, 2017). Model building followed a stepwise approach according to which fixed and random effects (by-speaker and by-rater) were progressively integrated into models and nested models were compared against one another by performing a Chi-squared test on their deviance statistics. The linear session predictor was centered on the

sample mean to reduce spurious slope-intercept correlations (Baayen, 2008), and a quadratic session predictor was computed by squaring the linear term to investigate curvature over time.

**Interrater Reliability and Rater Effects**

Ratings assigned to the native samples were first inspected to examine whether listeners reliably identified native speakers. One listener was removed from the data set because she assigned multiple native speakers a rating of 9, indicating that she had misinterpreted the directionality of the scale. No other anomalous ratings of the native data were observed. A continuous outcome variable was computed by aggregating ratings for the five sentences while maintaining the speaker, rater, and session factors. Interrater reliability was assessed separately for comprehensibility and accentedness using two two-way, agreement, average-measure intraclass correlation coefficients (ICC). For both constructs, the ICC was .86 (95% CI, .76 to .92). Mixed-effects models were fit to the ratings data integrating the following rater characteristics: English proficiency, percent daily Spanish use, frequency of interactions in Spanish with L2 Spanish speakers, familiarity with the characteristics of L2 Spanish, linguistics training, and teaching experience. No significant relationships emerged, indicating that raters assessed the clips similarly irrespective of their background and frequency of interaction with L2 Spanish speakers.

**Developmental Models**

Table 1 summarizes the taxonomy of developmental models fit. In these models, the linear session term examined linear growth over time, and the quadratic term examined curvature or quadratic growth over time. Both predictors were grand mean centered. Type refers to rating type, and interaction terms investigated whether comprehensibility and accentedness exhibited distinct rates of change over time. Random effects were included to account for by-speaker variability in development over time (e.g., adjusting the group-level coefficients slightly for each speaker), and by-rater variability in the evaluation of the speech samples.

An intercepts model was first fit including by-speaker and by-rater random intercepts (model 1) followed by two unconditional linear growth models, the first containing linear session as a fixed effect (model 2) and the second integrating by-speaker random slopes for the linear session predictor (model 3). Including quadratic session as a fixed effect significantly improved model fit (model 4), but the corresponding by-speaker random effect did not (model 5). The final unconditional growth model quantifying the effect of time on the ratings included fixed effects for linear and quadratic session, both of which were grand mean centered, by-speaker random slopes for linear session, and by-speaker and by-rater random intercepts.

Table 1

*Taxonomy of Models Fit.*

| | Model description | | | | | Test against prior model | |
|---|---|---|---|---|---|---|---|
| | Fixed effects | Random by-subject | Random by-rater | AIC | Δ AIC | Statistic | *p* |
| 1 | intercept | intercept | intercept | 15992 | | | |
| 2 | linear | | | 15892 | −100 | $\chi^2(1) = 102.38$ | < .001 |
| 3 | linear | linear | | 15870 | −22 | $\chi^2(2) = 24.48$ | < .001 |
| 4 | linear + quad. | linear | | 15848 | −23 | $\chi^2(1) = 24.72$ | < .001 |
| 5 | linear + quad. | linear + quad. | | 15846 | −2 | $\chi^2(3) = 7.48$ | .06 |
| 6 | model 4 + rating type (type) | linear | | 12800 | −3049 | $\chi^2(1) = 3049.28$ | < .001 |
| 7 | model 4 + type | linear + type | type | 9796 | −3604 | $\chi^2(5) = 3014.93$ | < .001 |
| 8 | model 4 + linear × type | linear + type | type | 9768 | −28 | $\chi^2(1) = 26.62$ | < .001 |
| 9 | model 4 + linear × type | linear × type | type | 9673 | −95 | $\chi^2(4) = 103.18$ | < .001 |
| 10 | model 4 + linear × type + quad. × type | linear × type | type | 9658 | −15 | $\chi^2(1) = 17.35$ | < .001 |

*Notes*. Fixed and random intercepts were included in all models. *Linear* and *quad. (quadratic)* refer to the treatment of session as a linear or quadratic predictor of ratings. The *linear* term tested linear growth over time, and the *quadratic* term examined curvature or quadratic growth over time. Both predictors were grand mean centered. *Type* refers to rating type. The interaction terms investigated whether comprehensibility and accentedness exhibited distinct rates of change (*linear × type*) and curvature (*quad. × type)* over time.

To examine whether comprehensibility and accentedness displayed distinct developmental trajectories, the effect of rating type (nominal, two levels; comprehensibility was set as the baseline against which accentedness was compared) was modeled following a similar stepwise procedure. The first conditional model included rating type as a fixed effect (model 6) and then as a by-speaker and by-rater random effect (model 7). A linear session × type interaction was subsequently integrated as a fixed effect (model 8) and as a by-speaker random effect (model 9). The final model (model 10, summarized in Table 2) incorporated the quadratic session × type interaction as a fixed effect.

Because comprehensibility was set as the baseline level of the rating type factor, the intercept and linear and quadratic session parameters refer to comprehensibility, whereas coefficients for model parameters involving rating type represent intercepts and trajectories for accentedness when combined with the corresponding (comprehensibility) baseline term. According to model estimates, at the outset of the study, participants began with average comprehensibility and accentedness ratings (i.e., intercepts) of 2.97 and 5.82 (intercept + rating type: 2.99 + 2.83 = 5.82). The negative coefficient for the linear session term (estimate = $-1.14$, $SE = .10$, $t = -11.91$) indicates that participants became more comprehensible and less accented over time. However, the positive coefficient for the linear slope by rating type interaction term (estimate = .54, $SE = .11$, $t = 4.77$) suggests that accentedness improved more gradually, at a rate of approximately $-.60$ units per session (linear session + linear session × rating type: $-1.14 + .54 = -.60$). The positive coefficient for the quadratic slope term (estimate = .13, $SE = .01$, $t = 11.20$) demonstrates that rate of change decreased over time. In other words, because negative coefficients index improvement given the lower-is-better operationalization of the rating scales in this study, the positive coefficient for the quadratic session predictor acts to constrain the magnitude of the negative growth trajectory observed for the linear term. The fact that the quadratic session by rating type interaction was also negative (estimate = $-.07$, $SE = .02$, $t = -4.18$) shows that the flattening out of the developmental trajectory observed for comprehensibility was not as pronounced for accentedness. Finally, the statistically significant by-speaker random effect for the linear session × type term suggests that participants displayed unique trajectories in each area over time. Figure 1 displays group trajectories as thick lines and model-estimated individual trajectories as thin lines. As is evident, comprehensibility displayed a steeper initial slope that began to level off significantly by the third session. In contrast, accentedness improved at a slower rate but exhibited less curvature over time. Figure 2 plots individual trajectories in more detail to demonstrate that whereas for some individuals comprehensibility and accentedness trajectories ran parallel, suggesting similar rates of change in both areas (e.g., 4, 22, 23), for other learners, comprehensibility either outpaced accentedness or accentedness did not change much at all (e.g., 9 and 24).

Table 2

*Final Model of L2 Ratings Data.*

| | Fixed effects | | | Random effects | |
| | | | | By Speaker | By Rater |
| *Parameters* | Estimate | *SE* | *t* | *SD* | *SD* |
|---|---|---|---|---|---|
| Intercept | 2.99 | .34 | 8.72 | .57 | 1.33 |
| Linear session | −1.14 | .10 | −11.91 | .31 | |
| Quadratic session | .13 | .01 | 11.20 | | |
| (Rating) Type | 2.83 | .47 | 5.96 | .49 | 1.91 |
| Linear × Type | .54 | .11 | 4.77 | .23 | |
| Quadratic × Type | −.07 | .02 | −4.18 | | |
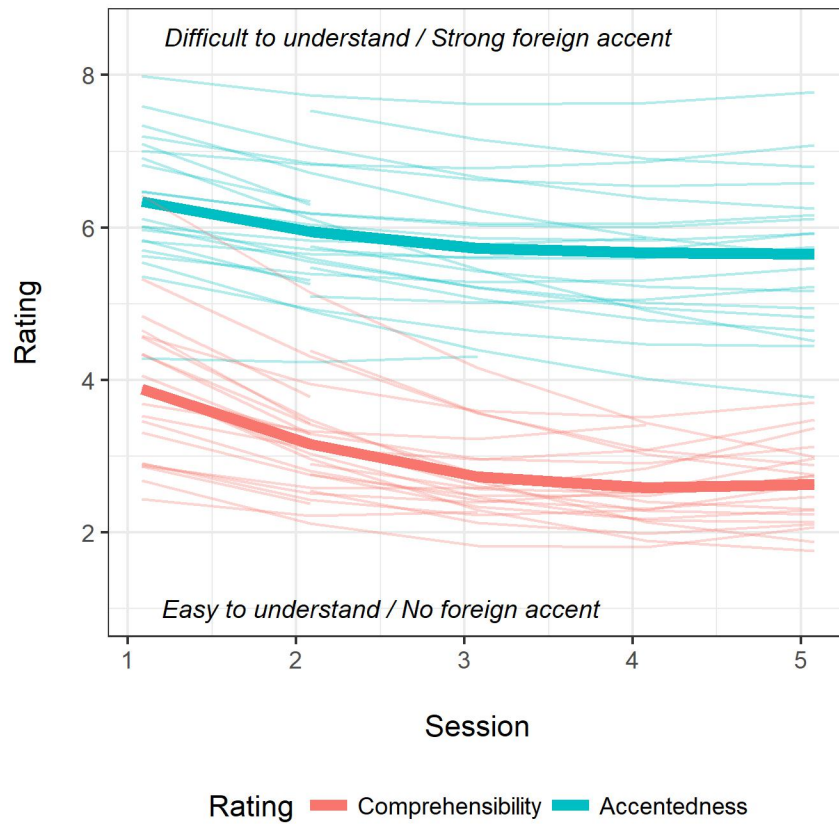
*Figure 1.* Development of L2 Comprehensibility and Accentedness over Five Sessions. Bold lines refer to the group trajectory and thin lines to individual trajectories.
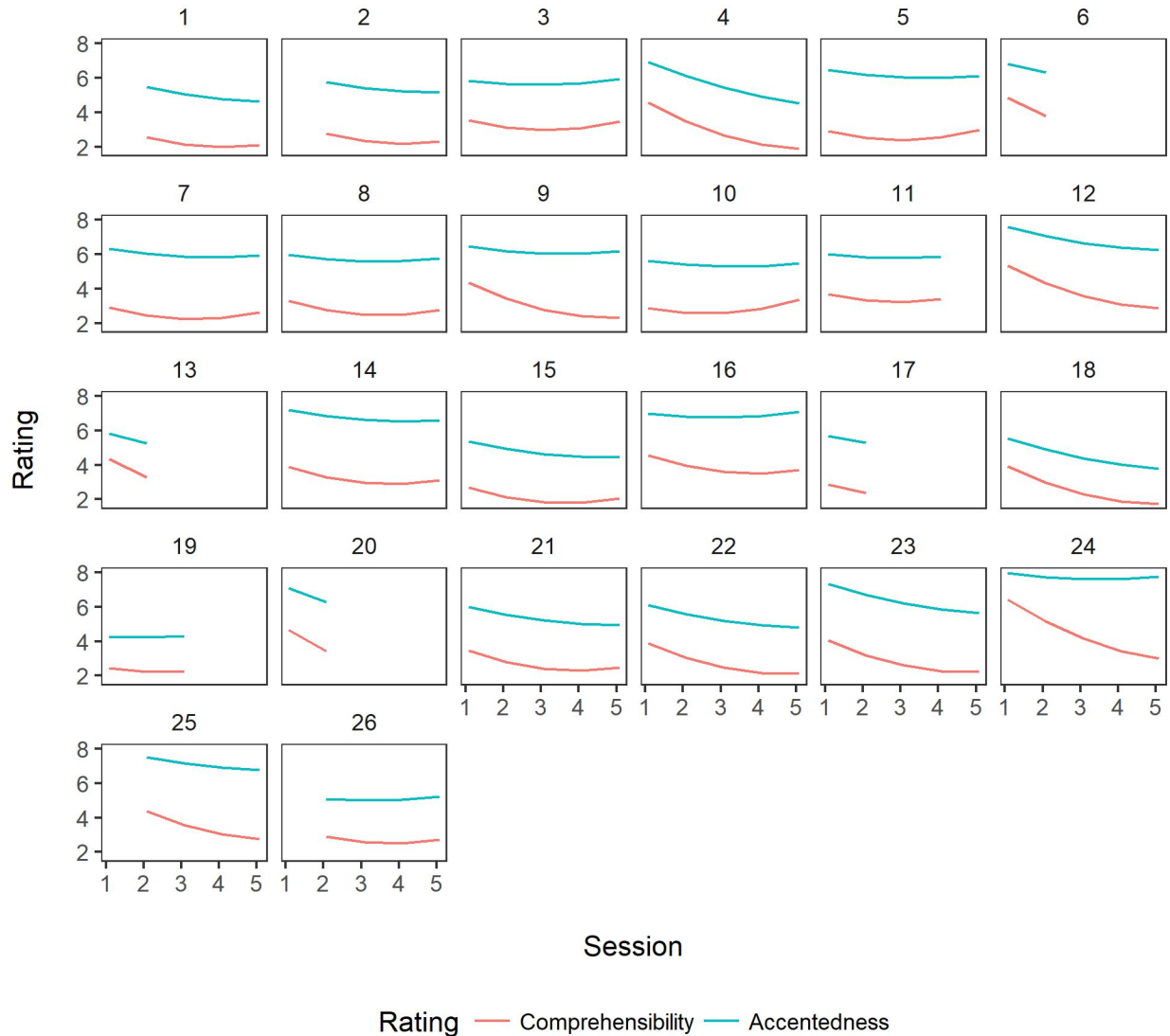
*Figure 2.* Individual Trajectories for L2 Comprehensibility and Accentedness over Five Sessions.

**DISCUSSION**

This study investigated novice L2 Spanish learners' comprehensibility and accentedness over a yearlong period encompassing data points distributed throughout their second, third, and fourth semesters of college-level Spanish language instruction. Mixed-effects models including linear and quadratic slopes were fit to examine rate of change and curvature over time in each dimension of L2 speech. At the outset of the study, learners were rated as comprehensible despite the presence of a moderate to strong foreign accent. Both comprehensibility and accentedness improved significantly while exhibiting differential rates of change. Comprehensibility improved more rapidly but was also subject to greater flattening over time, such that by the third data point the slope approached zero. On the other hand, even though

accentedness did not improve as rapidly over the first two sessions, it did not experience the same degree of leveling.

From a theoretical perspective, these findings intersect with Derwing and Munro's (2013) study on L2 immigrants living in Canada. In both cases, L2 speakers' pronunciation improved significantly with greater gains observed at the outset of the study, reinforcing the notion of a window of maximal opportunity for L2 pronunciation learning (Derwing & Munro, 2015). For classroom learners, the window of opportunity seems to extend over the first few semesters of language coursework, after which additional, targeted opportunities might be necessary to sustain development, particularly as far as accentedness is concerned. Results also intersect with Huensch and Tracy-Ventura's (2017) study tracking the development of various aspects of advanced L2 Spanish speakers' fluency while abroad. If we construct a continuum out of the speakers and observation periods contained within these three studies, then the tentative picture that emerges is one in which pronunciation development is characterized by short bursts of rapid change followed by enduring stretches of stabilization. In particular, precipitous shifts seem to occur at the onset of more intensive L2 contact, either via intensive communicative language training at university, as a result of an academic semester or year abroad, or due to relocation to an L2 environment. That is not to say that the speakers and contexts should be conflated, but rather that all three studies suggest similar developmental trajectories that may be additionally catalyzed or constrained by individual differences. However, more work is needed before a more definitive conclusion can be reached, particularly in studies tracking learners over even longer periods that encompass multiple watershed moments.

From a pedagogical perspective, results indicate that communicative language instruction can facilitate the development of more comprehensible and less accented speech in the language classroom even in the absence of pronunciation instruction. Given that accentedness improved at a slower rate, it could be beneficial to provide learners with additional targeted training to reduce accentedness. At the same time, significant individual differences in intercepts and rates of change were evident despite the fact that the 26 participants included in this study were all late learners of Spanish who had been placed into the same language course. Because the range of observed trajectories cannot be attributed to learners' previous experience or instructional factors, both of which were relatively uniform, future research should concentrate on investigating the role cognitive and attitudinal individual differences may play in the classroom context.

**ABOUT THE AUTHOR**

Charles Nagle (cnagle@iastate.edu) is an Assistant Professor of Spanish in the Department of World Languages and Cultures at Iowa State University. His research focuses on modeling the factors that underlie pronunciation development in L2 Spanish. His research interests include the perception-production link, individual differences in aptitude and motivation and their relationship to pronunciation learning, and the application of mixed-effects models to large data sets.

Iowa State University
World Languages and Cultures
3102 Pearson Hall
505 Morrill Road

Ames, Iowa 50011, USA
Phone: 515 294-6089

**REFERENCES**

Baayen, H. R. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. New York: Cambridge University Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). _lme4: Linear mixed-effects models using Eigen and S4_. R package version 1.1.-7. Retrieved from http://CRAN.R-project.org/package=lme4

Cunnings, I., & Finlayson, I. (2015). Mixed effects modeling and longitudinal data analysis. In L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research* (pp. 159-181). New York: Routledge.

Derwing, T. M., & Munro, M. J. (2013). The development of L2 oral language skills in two L1 groups: A 7-year study. *Language Learning, 63*(2), 163-185. doi:10.1111/lang.12000

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-based perspectives for L2 teaching and research*. Philadelphia, PA: John Benjamins.

Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*(3), 359-380. doi:10.1093/applin/amm041

Goldberg, D., Looney, D., & Lusin, N. (2015). Enrollments in languages other than English in United States institutions of higher education, Fall 2013. Retrieved from https://apps.mla.org/pdf/2013_enrollment_survey.pdf

Huensch, A., & Tracy-Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *The Modern Language Journal, 101*(2), 275-293. doi:10.1111/modl.12395

Kennedy, S., Foote, J. A., & Dos Santos Buss, L. K. (2015). Second language speakers at university: Longitudinal development and rater behaviour. *TESOL Quarterly, 49*(1), 199-209. doi:10.1002/tesq.212

Linck, J. A., & Cunnings, I. (2015). The utility and application of mixed-effects models in second language research. *Language Learning, 65*(S1), 185-207. doi:10.1111/lang.12117

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73-97. doi:10.1111/j.1467-1770.1995.tb00963.x

O'Brien, M. G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning, 64*(4), 715-748. doi:10.1111/lang.12082

R Core Team. (2017). R: A language and environment for statistical computing (Version 3.4.2). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439-462. doi:10.1093/applin/amv047

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2017). Lexical profiles of comprehensible second language speech: The role of appropriateness, fluency, variation, sophistication, abstractness, and sense relations. *Studies in Second Language Acquisition, 38*(4), 677-701. doi:10.1017/S0272263115000297

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*(4), 905-916. doi:10.1017/S1366728912000168