ISSN: 2707-1642

ORAL L2 COMPREHENSIBILITY AMONG ENGLISH IMMERSION LEARNERS WHEN GIVING SCHOOL SPEECHES

Comprensibilidad oral en la segunda lengua entre alumnos de inmersión en inglés al presentar discursos escolares

Dr. Susan K. Spezzini[4] *Dr. Rebecca L. Oxford*[5]

Abstract

This quantitative study examines the second language (L2) oral comprehensibility of 30English immersion learners, 15 girls and 15 boys, at an international school in South America (ISSA). At the end of 12th grade, these predominantly first language (L1) Spanish speakers gave speeches in English to a school audience. These speeches were recorded, and two audio extracts per student were selected. Native speakers of English listened to theses amples and rated them based on their own perceptions of each student‘s oral comprehensibility, i.e., *ease of understanding* (Derwing& Munro, 2005; Munro & Derwing, 2015). Female students received significantly higher comprehensibility scores than did male students. Students who had transferred to ISSA from other schools also tended to receive higher comprehensibility scores than those who had been schoole dat ISSA since kindergarten. The current study extends the field of comprehensibility research to the population of immersion learner sand to the curriculum component of school speeches as well as to the relationship between rater-perceived L2 comprehensibility and each student‘s gender and schooling. This study has direct implications for L2 instruction, assessment, and research.

Resumen

Esta investigación cuantitativa examina la comprensibilidad oral en la segunda lengua (L2) de 30 alumnos, 15 mujeres y 15 varones, que habían aprendido inglés por inmersión en una escuela internacional en América del Sur (ISSA, por sus siglas en inglés). Al final del grado 12, estos alumnos en su mayoría hablantes del español como primera lengua (L1) presentaron discursos en inglés frente a un público escolar. Estos discursos fueron grabados y dos fragmentos de audio por alumno fueron seleccionados. Hablantes L1 del inglés escucharon estas muestras y

[4] University of Alabama at Birmingham (USA), School of Education, spezzini@uab.edu
[5] University of Alabama at Birmingham (USA), School of Education, roxford@uab.edu

DOI: https://doi.org/10.47133/nemityra010105

las evaluaron de acuerdo a como percibieron la comprensibilidad oral, i.e., *facilidad para ser entendido* (Derwing& Munro, 2005; Munro &Derwing, 2015), de cada alumno. Las alumnas femeninas recibieron calificaciones de comprensibilidad significativamente superiores a las calificaciones recibidas por los alumnos masculinos. Los alumnos que habían sido transferidos a ISSA de otras escuelas también tendieron a recibir calificaciones de comprensibilidad superiores en comparación a aquellos cuyos estudios ocurrieron solamente en ISSA desde kindergarten. La actual investigación amplía el campo de estudios en comprensibilidad L2 a la población de alumnos L2 por inmersión y a la actividad curricular de discursos escolares como así también a la relación entre la comprensibilidad L2 oral con el género y el antecedente escolar de cada alumno. Esta investigación conlleva implicancias directas para la enseñanza, la evaluación y la investigación del L2.

Keywords: comprehensibility- intelligibility- pronunciation- gender- immersion education- international schools

Introduction

This quantitative study examines the second language (L2) oral comprehensibility of 30 English immersion learners, 15 girls and 15 boys, at an international school in South America (ISSA). At the end of 12th grade, these predominantly first language (L1) Spanish speakers gave speeches in English to a school audience. Excerpts from these school speeches were rated by native English speakers for perceived oral comprehensibility, i.e., *ease of understanding* (Derwing& Munro, 2005; Munro & Derwing, 2015). The current study extends the field of comprehensibility research to immersion learners at a school event and also to the relationship between immersion learners'oral L2comprehensibility and their gender and schooling. This study was guided by four research questions:

1) What is the general picture of oral L2 comprehensibility among English immersion learners in the 12th grade at an international school in South America?
2) How does the gender of these immersion learners relate to rater-perceived comprehensibility?
3) How does the schooling of these immersion learners relate to rater-perceived comprehensibility?
4) How do these learners' gender and schooling interact with respect to oral L2comprehensibility?

Literature Review
Comprehensibility and Intelligibility

Comprehensibility and intelligibility are rater-perceived constructs for assessing oral

L2 output. Based on scalar ratings, comprehensibility is the ―ease or difficulty with which a listener understands L2-accented speech‖ (Derwing, Munro, & Thomson, 2008, p. 360). Based on word transcriptions, intelligibility is ―the extent to which a listener actually understands an utterance‖ (Derwing& Munro, 2005, p. 385). Accentednesscontrasts with comprehensibility and intelligibility by presupposing a target language norm needed for rating ―how different a speaker‘s accent is from that of the L1 community‖ (p. 385). Varonis and Glass (1982) hypothesized overall comprehensibility as a major factor when judging pronunciation. Munro and Derwing (1995, 2001, 2015) and Derwing and Munro (1997, 2005, 2010) consistently showed that, although an L2 speaker might have a strong foreign accent, accentedness did not necessarily interfere with comprehensibility and intelligibility. These ―highly undisputed studies‖ (Rajadurai, 2007, p. 92) have been supported by more recent scholarship (Bergeron &Trofimovich, 2017; Crowther, Trofimovich, &Issacs, 2016; Saito, Trofimovich, &Issacs, 2016).

Higher ratings of oral L2 output were given by raters who shared the speakers‘ L1 (Smith &Bisazza, 1982); however, this advantage was also seen as small and inconsistent (Major, Fitzmaurice, Bunta, &Balasubramanian, 2002). For example, some studiesshowed similar ratings of L2 speech given by raters of diverse L1s (Munro &Derwing, 1995; Munro, Derwing, & Morton, 2006; Saito &Shintani, 2016). Similar findings resulted from comprehensibility studies of target languages other than English (Derwing, Thomson, & Munro, 2006; O‘Brien, 2014). Other studies showed similar ratings given by raters trained in linguistics and those untrained except for a brief explanation (Derwing, Rossiter, Munro, & Thomson, 2004). In short, raters did not seem to need linguistic training since they were measuring comprehensibility against their own perceptions regarding ease of understanding.

Comprehensibility and intelligibility may or may not improve substantially relative to a speaker‘s length of residence in an L2 context, as shown by Derwing and Munro (2010). Yet, when explicit pronunciation training is provided, comprehensibility and intelligibility can improve (Parlak, 2010). Other aspects of comprehensibility examined in the literature have included the effects of learner background on comprehensibility ratings (Crowther, Trofimovich, Saito, &Issacs, 2015), learner awareness of their own comprehensibility (Strachan, Kennedy,

&Trofimovich,

2019), linguistic influences of L2 output on listeners‘ comprehensibility ratings (Issacs&Trofimovich, 2012), listeners‘ familiarity with speakers‘ L2 accent (Winke&Gass, 2013), and listeners‘ beliefs in what underlies their judgments (Hayes-Harb& Hacking, 2015; Jun & Li, 2010). Earlier comprehensibility studies often involved L2 performance tasks, in which participants (usually adults) responded to researcher-manipulated prompts in

laboratories or laboratory-like settings (i.e., secluded rooms). Rajadurai (2007) criticized such perception studies of oral L2 output for their (a) reliance on laboratory-based, artificial, decontextualized language; (b) inadequate consideration of interactional contexts; and (c) emphasis on the Inner Circle of World Englishes (Kachru& Nelson, 1996). She suggested a more qualitative, interpretive, and socioculturally realistic dimension and asserted that a need existed for moving the methodology to more naturalistic venues, such as school settings. These organic research settings are important because comprehensibility and intelligibility are linked to the linguistic and sociocultural context, not just to the speaker and listener (Gumperz, 1992; Kennedy &Trofimovich, 2008; Pickering, 2006). Our study of ISSA students seeks to address this restriction by expanding the field of comprehensibility studies to school-aged learners at a naturally-occurring school event, that of school speeches.

Gender Differences in Language Use

Gender has repeatedly been reported as influencing L2 learning and use (Brantmeier, Schueller, Wilde, &Kinginger, 2007; Brown, 2006; Ellis, 1994; Motallebzadeh&Nematizadeh, 2011; Oxford, 1995; Pavlenko, 2008; van der Slik, van Hout, &Schepens, 2015; Zhu, 2009). For example, women are more likely to use standard and prestigious speech forms, and men ―orient their speech more to localized norms‖ as a ―phonological marker of identity‖ (Pennington, 1996, p. 5, 16). Men have also received ―statistically less native ratings of accentedness than females‖ (Munro & Mann, 2005, p. 329). Additionally, Major (2004) has shown gender-differentiated stylistic variation in L2 phonology. Issues related to gendered L2 use have also been examined through the theoretical lens of investment and imagined communities (Norton, 1997; Norton &Pavlenko, 2004). To enhance further understanding of gendered L2 variants, Moyer (2010) called for additional research with the ultimate goal to help language learners—specifically males—in reaching higher levels of L2 attainment.

Originally, gender variability in language use had been explained by social

psychologists via theories on covert prestige and accommodation. *Covert prestige* is the attribution of prestige to a nonstandard variant that is both hidden and divergent from accepted social values (Trudgill, 2000). In a U.S. urban setting, Labov (1970) found gender-differentiated, standard and non-standard L1 use beginning as early as age 10, when children ―come under the full influence of the preadolescent peer group‖ (p. 34). *Accommodation* is ―a multiply-organized and contextually complex set of alternatives, regularly available to communicators in face-to-face talk. It can function to index and achieve solidarity with or dissociation from a conversational partner, reciprocally and dynamically‖ (Giles &Coupland, 1991, pp. 60-61). The helpful theories of covert prestige and accommodation established a basis for examining convergence, divergence, and/or maintenance of linguistic features in the language varieties of reference groups in our study.

The general consensus is that sex differences are biological and gender differences are socioculturally constructed (Schneider, Gruman, & Coutts, 2005). AlthoughTannen (1990) describes men‘s speech as more competitive and women‘s speech as more cooperative, Eckert and McConnell-Ginet (2003) contend that such generalized statements are insensitive to differences in sociocultural contexts. Their mention of masculine vernacular styles, which are related to ―positioning‖ within certain sociocultural contexts, is relevant to our study.

Gendered L2 Use among Immersion Learners

In a study of language use among immersion middle school students, gender was ―more significant than L1 in overall language output‖ (Potowski, 2007, p. 114). At a Spanish-English dual language school, girls positioned themselves as more compliant and more likely to follow classroom rules, which included the use of L2. The girls‘ investment in their own academic standing was coupled with activities beyond the classroom that were also associated with L2 use. The boys were seen to use the L2 less frequently and less accurately.

In a study of a high school French immersion setting, students used sociolinguistic variants based on gender, social background, syntactic context, languages spoken at home, and extracurricular exposure to native L1 use (Mougeon&Rehner, 2001). This specific study suggested that ―female students show a stronger tendency than do their male counterparts to prefer the variants that are part of standard usage and that are likely to have been favored by their teachers‖ (p. 408).

Swain and Johnson (1997) asked whether ―the non-target-like proficiency of immersion students [is] an inevitable outcome of immersion‖ programs and whether ―pedagogical practices might improve the attainment of target-like

proficiency‖ (p. 15). To that end, several studies examined L2 attainment at immersion schools in Canada and the U.S. (Potowski, 2007; Tarone& Swain, 1995) and at international schools in Europe, Asia, and South America (Carder, 2007; de Courcy, 2002; de Mejìa, 1998; Spezzini, 2004). Even though immersion learners at a given school may experience almost identical L2 opportunities and receive similar comprehensible input (Krashen & Terrell, 1983), they do not necessarily produce similar comprehensible output (Swain, 1985). Hence, within the same context, different individual investments seem to produce different L2 outcomes (Norton, 1997).

These immersion studies produced findings that are supported by the theoretical framework in gender discussed earlier: girls more closely approximate L2 target forms. Situated in naturalistic settings, these studies employed relatively traditional approaches for conducting classroom-based research. In contrast, our study, by using rater-perceived comprehensibility to examine immersion learners‘ oral L2 output at an international school, adds a new dimension to studies of L2 oral comprehensibility.

Methodology

Setting

Our study took place at ISSA, an American overseas school located in South America. Founded in the 1950s, ISSA is a private nonsectarian K-12 school that serves dependents from embassies and multinational firms as well as dependents of local families with the means to pay high tuition. ISSA is one of over 850 English-medium schools in countries with official languages other than English (Baker &Jones, 1998). As such, ISSA fits the definition of an international school (Hayden, Levy, & Thompson, 2015; Johnson & Swain, 1997). Because ISSA provides instruction in two languages, albeit with unequal distribution, its curriculum can be defined as bilingual education (Abello-Contesse, Chandler, Lopez-Jimenez, & Chacon-Beltran, 2013; Baker & Jones, 1998; de Mejia, 2005).

ISSA students are schooled almost entirely in English with two classes (*Castellano* and *Estudios Sociales*) in Spanish. WithinISSA‘s institutionally-defined speech community, students use mainly English for academic purposes and Spanish for social purposes (Spezzini, 2004). Based on such task-specific discourse by ISSA students, a student who had participated in an earlier study explained how she and her
classmates ―do not talk [English] at all as the people do in the U.S.‖ (p. 421).

Because most ISSA students have a home language other than English but are schooled mainly in English, their schooling can be defined as English immersion. To that end, ISSA embodies six descriptors of immersion education (Swain & Johnson, 1997): L2 as a medium of content instruction, overt support for the L1, program goal of additive bilingualism, L2 exposure confined mainly to the classroom, similar (and limited) L2 proficiency upon admission (usually kindergarten or pre-kindergarten), and school culture that reflects the local culture. Also definable as foreign language immersion in an international context (Lyster, 1999), ISSA represents a type of ―hybrid‖ situation (Green & Oxford, 1995). ISSA's academic environment resembles an English as a second language (ESL) setting where students use English (L2) in the classroom with teachers who are native speakers of English. Its social environment resembles an English as a foreign language (EFL) setting where almost all students use Spanish (L1) outside of the classroom, both on campus and in the community.

Participants[1]

Study participants included 30 ISSA students, 15 girls and 15 boys. They participated in this study as 12th graders during the last month of their final year of schooling. Their mean age was 18 years and10 months. Their home languages were 25 Spanish (from 4 countries), 3 Chinese, 1 Korean, and 1 Portuguese. With respect to schooling, 14 students had been at ISSA since kindergarten (or pre-kindergarten), and 16 had transferred from other schools to ISSA between first grade and tenth grade.

Procedure

We considered three already-tested constructs to measure oral L2 output: *accentedness*, *intelligibility*, and *comprehensibility*. As explained by Munro and Derwing (2015) and Levis (2018), these constructs are related yet different. *Accentedness* evaluates a speaker's foreign accent by having listeners provide scalar ratings of the distance between L2 output and an idealized native-speaker norm. *Intelligibility* evaluates a speaker'sability to be understood by having listeners do a word-by-word transcription. *Comprehensibility evaluates a* speaker's ease in being understood by having listeners provide scalar ratings on their perception of L2 output as being easy or difficult to understand. We rejected *accentedness* as the construct for our study because it might imply a negative view of L2 output. We rejected *intelligibility* because it would require transcriptions rather than scalar ratings. We selected *comprehensibility* as our construct because its scalar ratings would facilitate data collection and analysis and also because it fit conceptually to our study.

After choosing comprehensibility as our dependent variable, we created a Likert-scale instrument modeled after one used by Munro and Derwing (1995). Based on an assumption upheld by Southwood and Flege (1999), this instrument partitioned listener perceptions of oral output into equal intervals. With a similar instrument, Derwing et al. (2004) obtained acceptable interrater reliability values in the .70s using a procedure recommended by Hatch and Lazaraton (1991)—computing ―mean interrater correlations after conversion to Fisher Z scores‖ (Munro &Derwing, 2001, p. 459). Reliability values for the current study are provided in the next section.

We also wrote instructions modeled after the same study (Munro &Derwing, 1995). Raters were to listen carefully to each speech sample, determine ease in understanding the speaker, and, during a 2-second pause, circle a rating from 1 (extremely easy to understand) to 9 (impossible to understand). Raters were instructed to judge their own ease in understanding each sample and not whether they had actually understood. Although our instrument and instructions were modeled on earlier studies, our study differed by using L2 speech samples produced by immersion learners when giving school speeches as a required curricular component during a naturally-occurring school event.

We piloted an initial 2-page version of our instrument by asking ISSA teachers to rate recorded samples from these speeches. This pilot study took place immediately following a faculty meeting. Based on raters‘ feedback and a broad dispersion in their ratings, we reduced the scale from 9 to 7and expanded the instrument from two to three pages. In this revised instrument, the first page provided general instructions and a warm-up, the second page was for circling the ratings of Set A samples, and the third page for circling the ratings of Set B samples. We tested this revised 3-page instrument with two teachers. Their responses confirmed the use of this revised instrument for our study.

Raters[1]

Our study involved 26 raters: 12 piloted the initial instrument, 2 tested the revised instrument, and another 12 participated in the actual study. This total set of 26 raters, which included the final set of 12 raters, compares well to the 10, 18, 26, and 28 raters used in other studies (Crowther, Trofimovich, &Issacs, 2016; Derwing et al., 2004; Munro &Derwing, 1995; Munro, Derwing, & Morton, 2006). Except for a brief explanation of the rating procedure, our raters were untrained in rating L2 comprehensibility. They complemented our study‘s school-based setting by representing how real-life interlocutors might react to ISSA students during authentic interactions. Our use of untrained raters is supported by comprehensibility studies that showed strong correlations between responses

from untrained raters and those from phonetically-trained raters (Derwing et al., 2004).

Other researchers have used college students as raters because of their availability at the institutions housing the laboratory-based studies (Derwing et al., 2004). In contrast, we intentionally selected teachers as raters. They were familiar with school-aged learners; yet, as recent arrivals to ISSA, they were unfamiliar with the L2 output of ISSA students. Furthermore, because almost none of these teacher participants spoke Spanish, these teachers could not unconsciously adjust their ratings based on knowledge of a speaker‘s L1. Eager to hear voices of actual ISSA students, they focused intently on listening to the samples and rating their own perceptions.

Data Collection

Because ISSA students use English (L2) primarily for academic interactions and Spanish (L1) for social interactions (Spezzini, 2004), we decided to use the students‘ academic English from a school event to determine their oral L2 comprehensibility. We selected an annual pre-graduation event, that of 12th graders giving formal speeches. Over a two-day period, each 12th grader delivered a 15-minute speech to a school audience consisting of students from other grades, staff, and parents. We recorded these speeches with a commonly available recording device. The first author also observed all of the speeches and took notes. Students were dressed in business attire representative of the countries and agencies studied in their government class. First, they read speeches from scripts prepared in their English class. Then, they explained what they had learned from this project. Similar to how other studies have used both read passages and extemporaneous utterances (Derwing& Munro, 2005; Gass, Sorace, &Selinker, 1999), we used excerpts from each student‘s speech and post-speech commentary.

We grouped the selected excerpts into two listening sets. Set A, *Describing My Country,* contained samples from the actual speeches. Set B, *What I Learned,* contained samples from the post-speech commentaries. By combining each student‘s ratings from sample A and sample B, we obtained a comprehensibility score that more closely represented that student‘s range of school-based speaking.

Before rating the study samples in Sets A and B, the raters did a warm-up with 5non-study samples from the speeches and another 5 non-study samples from the commentaries. This 10-item warm-up served to see if raters were following instructions; it was not used to elicit rater feedback. Moreover, to ensure the raters‘ readiness to start, an additional warm-up item prefaced Set A and also Set B. These 12 warm-up items were not among the study-related speech samples.

Because our data collection took place during a school event, we recorded all 34 students in the12th grade who were giving speeches. Samples from all 34 were included in the listening task for raters. However, for data analysis purposes, we deleted the ratings of four students: two who could not be heard clearly because of a technological failure, a third who had not been in the government class and hence spoke on a different topic, and a fourth who was a native English speaker (NES)[2]. To the advantage of this study, the samples from this NES served as checks to ensure that raters had not lost their place, a control purpose used in other studies (Derwing et al., 2004; Munro &Derwing, 1995).

All speech samples contained a minimum of non-communicative pauses and were void of unfamiliar place names that might have affected a rater‘s perception of comprehensibility. The samples were ordered differently on the listening track from how the corresponding speeches had been delivered. Each sample was preceded by that student‘s study-assigned number spoken by the first author.

The first part of the listening was 22 minutes long. It contained oral instructions (including the 10-item warm-up) and Set A samples, which ranged from 12 to 15 seconds in length. The second part was 24 minutes long and contained Set B samples, which ranged from 25 to 30 seconds in length. The combined samples ranged in length from 12 to 30 seconds, which was similar to the length, from 7 to 30 seconds, in other comprehensibility studies (Derwing et al., 2004; Munro et al., 2006). As in other comprehensibility studies (Munro &Derwing, 1995), we also selected samples that began and ended at normal breaks in the utterances. Because of inherent factors related to giving formal speeches (e.g., use of visual aids, repeated pauses, nervousness), the Set A samples were relatively short. This was counterbalanced with longer samples in Set B.

The comprehensibility ratings were obtained from 12 ISSA teachers on the second day of a weeklong faculty development program for recently-arrived staff. This rating task was the first activity during a morning-long session on second language acquisition. To optimize this listening experience, we conducted the ratings in the same room and with the same equipment used at ISSA for administering Advanced Placement language exams.

The entire rating activity lasted 7 minutes for instructions and warm-up, 15 minutes for Set A, and 24 minutes for Set B. This 46-minute duration is about midway between the durations infrequently-cited comprehensibility studies: 35 minutes plus brief familiarization (Munro et al., 2006) and 60 minutes with one short break (Derwing et al., 2004). Our 68 samples (34 in Set A and 34 in Set B) were slightly more than the 63 samples used by Derwing et al. Our raters judged 68 samples, which included the 60 samples used in this study (2 from each of 30 students included) plus an additional 8 samples not used in this study (2 from each

of 4 students excluded).

Each of our 12 raters assigned two ratings to each student—one rating for Set A and another for Set B. This produced a total of 24 ratings per student (i.e., 12 raters x 2 ratings). The mean of these 24 ratings became each student‘s rater-perceived L2 comprehensibility score. This score quantified a student‘s spoken English based on how NES raters perceived their own ease in understanding.

Although our raters marked their perceptions on an instrument ranging from 1 (extremely easy to understand) to 7 (impossible to understand), we reversed the scale to facilitate meaningful correlation analyses for other parts of the greater study.[3] Through this reversal, low value ratings corresponded to low comprehensibility and high value ratings to high comprehensibility. In this article, all scores are from this reversed scale—with 1 for the lowest level of comprehensibility (impossible to understand) and 7 for the highest level of comprehensibility (extremely easy to understand).

Data Analysis

We performed all statistical analyses on SPSS® for Windows. To answer the first research question, we calculated descriptive statistics of means, standard deviations, and ranges for the scores of the dependent variable (comprehensibility) in relationship to the independent variables (gender and schooling). In response to the second and third research questions, which asked whether any main effect occurred because of gender and schooling, respectively, we employed two-way analysis of variance (ANOVA) and Tukey. The ANOVA also served to answer the fourth research question, which asked about the interaction between gender and schooling with respect to comprehensibility

To determine whether raters‘ scores were highly correlated to each other (interrater reliability), we calculated the Intra-Class Correlation Coefficient (ICC). Though we knew that raters‘ scores were not identical in absolute terms, the ICC provided useful information by showing if raters were consistent based on the similarity of their relative ratings.

Because our research design involved averaging multiple ratings for each item, we selected the ICC model called *average measure reliability*. By doing this, we also determined that our study met this model‘s criteria, that of having a reasonable number of raters (n=12) for forming a stable average. ICC‘s *average measure reliability*, which uses the mean of all ratings as the unit of analysis, provides the reliability of the mean of the ratings of all raters. ICC is a measure of homogeneity, one that approaches 1.0 when any given row (ratings for a given student) tends to have the same values for all columns (i.e., raters). The average measure reliability

for either two-way random effects or two-way mixed models is the same as Cronbach‘s alpha.

We obtained an ICC for Set A of alpha = .918 and an ICC for Set B of alpha = .942. These high ICC values show a high level of interrater reliability. The lower ICC value for Set A (.918) than for Set B (.942) was reasonable, because students were visibly more nervous during the formal speeches (Set A) than the post-speech commentaries (Set B). The interrater agreement indicated that students‘ L2 output (i.e., recorded speeches and commentaries) had provided the stimuli for the ratings. In other words, if linguistic features of the recorded stimuli were of greater importance in determining a comprehensibility score and, in turn, if individual-rater factors were of lesser importance, then we would expect that a high degree of consensus among listeners would lead to strong interrater agreement (Munro et al., 2006). This consensus among raters (i.e., interrater reliability) shows that our rating activity in an immersion education setting and its comprehensibility scores are reliable.

The high levels of interrater reliability also provided criteria for establishing the construct validity of comprehensibility in this immersion education setting. For their laboratory-based study, Derwing et al. (2004) considered ―an examination of reliability in listeners‘ judgments of fluency [as] important in establishing the construct validity of perceived fluency‖ (p. 658). They also found that ―comprehensibility was significantly correlated with fluency judgments‖ (p. 669). In a parallel manner, the interrater reliability for the listeners‘ judgments of comprehensibility in our study served to establish the construct validity of rater-perceived comprehensibility.

Despite having established construct validity, we felt that the 3:1 female-to-male ratio of our 12 raters dictated a need to check whether gender, of both raters and speakers, had influenced comprehensibility ratings. Although unbalanced gender ratios (7:3 and 22:6, female-to-male) had existed among raters in other comprehensibility studies (Derwing et al., 2004; Munro et al., 2006), examinations of possible gender-related bias had not been conducted. To determine whether our raters consistently gave higher scores to speakers of a specific gender, we performed a cross-gender comparison. On a 7-point scale, girls received a mean score of 6.0 from female raters and 5.6 from male raters, and boys received 4.9 from female raters and 4.5 from male raters. In other words, both girls and boys received mean ratings that were 0.4 higher from female raters than from male raters. Thus, female raters rated speakers more positively than did male raters and did so consistently for girls

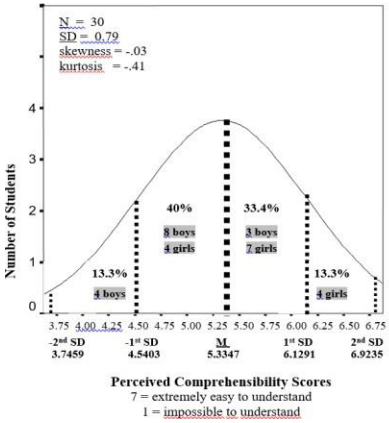
and boys. Perhaps female raters (or females in general), as compared with male raters (or males in general), try harder to understand others because of greater compassion or concern, built in through socialization. In short, the high ratings received by girls and the low ratings received by boys did not result from any gender-related bias. This cross-gender analysis indicated that raters of both genders found it easier to understand ISSA girls and more difficult to understand ISSA boys.

Results

Research Question 1: Overall Picture of Oral L2 Comprehensibility

An initial picture of oral L2 comprehensibility is provided in Figure 1. Here, the descriptive statistics of means and standard deviations from the students' comprehensibility scores represent a fairly normal distribution. On a scale from 1(lowest comprehensibility) to 7 (highest comprehensibility), the mean was 5.3347, with a standard deviation of 0.79 and skewness of -0.03. Of the 30 students, 14 placed above the mean and 16 below the mean. The 15thranked student had a score of 5.3333, which was just below the mean. Had her score been 0.0015 higher, it would have fallen above the mean leading to exactly half of the students (15) above the mean and also half (15) below the mean.

Figure 1. Gender Distribution of L2 Comprehensibility Scores



As shown in Figure 1, 73.4% fell between the positive and negative first standard deviations, with 33.4% above and 40% below. This is similar to a normal distribution of 68%, with 34% in each direction. Had the 15th-ranked score been 0.0015 higher, the distribution between the mean and first standard deviation

would have been even with 36.7% in each direction. The other 26.6% fell between the first and second standard deviations, exactly half above and half below, similar to the 27.3% in a normal distribution. No scores fell beyond the second standard deviation in either direction.

To better understand the normal curve in Figure 1, we superimposed gender by inserting shaded comments below the curve. Only 3 of the 15 boys scored above the mean, yet none above the first standard deviation. Similarly, only 4 of the 15 girls scored below the mean, yet none below the first standard deviation. This inverse score distribution suggested that girls had higher comprehensibility than boys, which was then corroborated by significance testing. Results showed that rater-perceived comprehensibility in this immersion education setting corresponded positively to earlier studies where girls outperformed boys in oral L2 output (Pennington, 1996; Swann, 1999).

This general picture of comprehensibility is expanded in Table 1 which provides the means and standard deviations for each gender as well as the range of scores for each gender. Table 1 shows how girls spoke with significantly higher comprehensibility (M=5.81) than did boys (M=4.86). Given a spread of 3.08 points between the lowest overall score (3.75) and the highest overall score (6.83), the 0.95 difference between the gender means is noteworthy.

Table 1. *Descriptive Statistics for Comprehensibility as Related to Gender*

Students	*N*	Minimum Score	Maximum Score	*1.1.1. M*	*SD*
Girls	15	4.75	6.83	5.81	0.61
Boys	15	3.75	6.08	4.86	0.68
TOTAL	**30**	**3.75**	**6.83**	**5.33**	**0.79**

Note: Each student's score was the mean of 24 ratings, which consisted of two ratings from each of the 12 raters. The ratings ranged from 1 (lowest comprehensibility = impossible to understand) to 7 (highest comprehensibility = extremely easy to understand).

To further expand this general picture of oral L2 comprehensibility, we examined the means, standard deviations, and score ranges for each gender vis-à-vis schooling. To examine the independent variable of schooling, students were divided into three groups:

KC = Entered ISSA in **Kindergarten** and were **Citizens** of the country where ISSA was located;

TN = **Transferred** to ISSA from another school and were **Not** citizens of the country where ISSA was located; and

TC = **Transferred** to ISSA from another school and were **Citizens** of the country where ISSA was located.

Table 2 illustrates the relationship between the dependent variable, comprehensibility, and both of the independent variables, gender and schooling. In the KC group, the lowest score among girls (4.75) was higher than the highest score among boys (4.71). In other words, each KC girl spoke with higher comprehensibility than did each KC boy. Table 2 also shows that the lowest group mean among girls (KC group: $\underline{M}$ = 5.68) was higher than the highest group mean among boys (TN group: $\underline{M}$ = 5.19). This table further suggests that within each gender, higher comprehensibility was attained by students who had transferred from other schools. Thus, schooling appears related to comprehensibility differences across both genders.

Table 2. *Descriptive Statistics for Comprehensibility as Related to Gender and Schooling*

Students	*N*	Minimum Score	Maximum Score	*1.1.1. M*	*SD*
			Girls		
KC[a] girls	9	4.75	6.67	5.68	0.52
TN[b] girls	4	5.04	6.83	5.98	0.92
TC[c] girls	2	5.88	6.21	6.04	0.24
			Boys		
KC[a] boys	5	3.75	4.71	4.23	0.45
TN[b] boys	4	4.67	6.08	5.19	0.63
TC[c] boys	6	4.46	6	5.17	0.54

Note: Each student's score was the mean of 24 ratings, which consisted of two ratings from each of 12 raters. Ratings ranged from 1 (lowest comprehensibility=impossible to understand) to 7 (highest comprehensibility=extremely easy to understand).

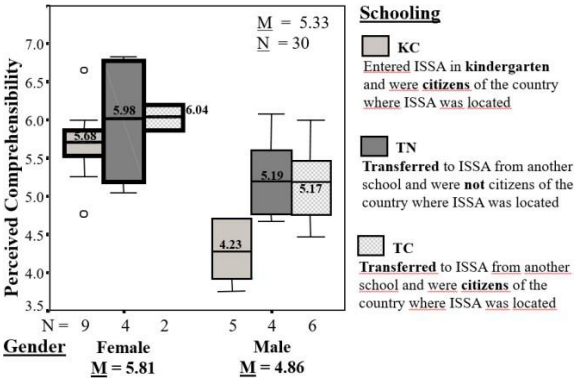
[a]KC = Entered ISSA in **Kindergarten** and were **Citizens** of the country where ISSA was located.

[b]TN = **Transferred** to ISSA from another school and were **Not** citizens of the country where ISSA was located.

[c]TC = **Transferred** to ISSA from another school and were **Citizens** of the country where ISSA was located.

The data from Table 2 is presented visually in Figure 2. The box plot illustrates how the three girl schooling groups (left side) have higher comprehensibility than the three boy schooling groups (right side). It also illustrates how the KC boys have much lower comprehensibility than both groups of transfer boys (TN and TC). This box plot strengthens the general picture of oral L2 comprehensibility among ISSA 12th graders and, by doing so, contributes towards answering Research Question #1.

Figure 2. *Gender and Schooling Distribution of L2 Comprehensibility Scores*



(Spezzini & Oxford, 2003)

Research Questions 2 and 3: Gender Effect and Schooling Effect

Analysis from a two-way ANOVA is presented in Table 3. Here, gender and schooling were the independent variables, and comprehensibility was the dependent variable. Regarding Research Question 2, results revealed a significant main effect for gender ($F(1, 24) = 19.12$, $p<.001$), with an effect size (partial $\eta^2 = .443$) that was small but not altogether distant from medium. Small effect sizes start approximately at .2, while medium effect sizes begin roughly at .5 (Aron, Aron, & Coups, 2005).

Regarding Research Question 3, results from the two-way ANOVA (Table 3) revealed a significant main effect for schooling ($F(2, 24) = 3.94$, $p<.05$), with an effect size (partial $\eta^2 = .247$) that was small. To identify the precise location of significance in all possible comparisons related to schooling (exclusively or nonexclusively at ISSA), Tukey's LSD test was conducted. This analysis is

provided in Table 4).

Table 3. *Factorial ANOVA for the Dependent Variable of Comprehensibility and the Independent Variables of Schooling (3 groups) and Gender (male/female): Tests of Between-Subjects Effects*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial ETA Squared
Corrected Model[a]	10.059	5	2.012	5.858	0.001	0.55
Intercept	705.5	1	705.5	2054.439	0	0.988
SCHOOLING[b]	2.707	2	1.353	3.941	0.033	0.247
GENDER	6.561	1	6.561	19.107	0	0.443
SCHOOLING* GENDER	0.65	2	0.325	0.946	0.402	0.073
Error	8.242	24	0.343			
Total	872.078	30				
Corrected Total	18.3	29				

[a]R^2 = .550 (Adjusted R^2 = .456)
[b]Schooling Groups:
KC=entered ISSA in kindergarten and were citizens of the country where ISSA was located
TN=transferred to ISSA from another school and were not citizens of the country where ISSA was located
TC=transferred to ISSA from another school and were citizens of the country where ISSA was located

Table 4. Pairwise Comparisons using Tukey's LSD for Dependent Variable Comprehensibility

Schooling Groups		Mean Difference	Standard Error	Sig.	95% Confidence Interval for Difference	
(I) group	(J) group	(I-J)			Lower Bound	Upper Bound
1 = KC group	2	-0.626	0.264	0.026	-1.171	-8.18E-02
	3	-0.647	0.29	0.035	-1.245	-4.93E-02
2 = TN group	1	0.626	0.264	0.026	8.18E-02	1.171
	3	-2.08E-02	0.316	0.948	-0.674	0.632
3 = TC group	1	0.647	0.29	0.035	4.93E-02	1.245
	2	2.08E-02	0.316	0.948	-0.632	0.674

Based on estimated marginal means.
*The mean difference is significant at the .05 level.

Adjustment for multiple comparisons: Least Significant Difference (equivalent to no adjustments).
Results revealed that differences between means were statistically significant at the $p<0.05$ level between the KC group and both the TN and TC groups ($p=.026$ and $p=.035$, respectively). In other words, non-transfer students, the KC group (exclusively at ISSA), were significantly different from transfer students, the TN and
TC groups (nonexclusively at ISSA). The 16 transfer students (8 TN and 8 TC)

exhibited higher oral L2 comprehensibility than the 14 non-transfer students (KC group). This was not surprising since, before coming to ISSA, many transfer students had been attending schools where English was their social language.

This post hoc test also showed no significant difference in oral L2 comprehensibility between the TN group (transfer students who were not citizens of the country where ISSA was located) and TC group (transfer students who were citizens of the country where ISSA was located). Both of these transfer groups (TN and TC) were somewhat similar to each other in oral L2 comprehensibility, and both were significantly different from the non-transfer group (KC). This result

from our study, which was based on the application of rater-perceived comprehensibility in an international school's immersion education setting, conforms with other studies that showed less native-like oral L2 development among immersion learners when restricted to immersion education settings in the United States and Canada (Cloud, Genesee, &Hamayan, 2000) and more native-like oral L2 development when access is provided to additional L2 use, especially beyond academic settings (Lyster, 1999; Mougeon&Rehner, 2001).

The effect size findings suggested a somewhat greater impact of gender than schooling on ISSA students' comprehensibility. While both the effect sizes were in the small category, the effect size for gender was closer to medium than was the effect size for schooling.

Research Question 4: Interaction Effect

For Research Question 4, the interaction of both independent variables, gender and schooling, on the dependent variable, comprehensibility, was examined. Results from the two-way ANOVA (Table 3) showed no significant interaction between gender and schooling (F (2, 24) = .95, p = .402, partial η^2 = .073) among the three schooling groups (KC, TN, and TC). This finding indicated that it was not possible to reject the null hypothesis (H_o = There is no significant interaction between gender and schooling for the dependent variable). Consequently, although the responses to Research Questions 2 and 3 had shown that gender and schooling each had a significant impact individually on perceived oral L2 comprehensibility, similar significance could not be established regarding the interaction of these two independent variables on comprehensibility.

Discussion

This section evaluates the use of a context-sensitive research methodology for establishing oral L2 comprehensibility. It discusses insights from this study; cites this study's relationship to other studies; explores potential factors for observed gender difference in comprehensibility; and examines the nexus of schooling, L2 proficiency, and language learning success.

Context-Sensitive Methodology

The current study responds to Rajadurai's (2007) call for a ―reconceptualized, context-sensitive view‖ of intelligibility and comprehensibility, one that recognizes ―the dynamics of context … and the legitimacy of varieties of English‖ (p. 96). Because findings ―are always relative to the conditions under which they came

about, always shaped to some extent by the research questions asked, and always of necessity partial‖ (p. 96), we guard against direct application of these findings to other settings. However, we believe that extrapolation to immersion education settings, such as other relatively closed speech communities with institutionally-defined memberships, could be warranted.

By having untrained listeners rate the oral L2 comprehensibility of immersion learners at an international school, this study has extended the construct of comprehensibility to a new population and a new setting. Instead of being used mainly in researcher-manipulated settings, the venue for comprehensibility studies has now been extended to the curricular component of student speeches at a naturally-occurring school event. Moreover, comprehensibility, a construct usually used with adult learners, has been shown to be equally as valid for use with school-aged learners.

Multiple Insights and Relationship to Other Studies

This study also provides multiple insights to oral L2 comprehensibility among immersion learners at an international school, seen here as an institutionally-defined closed speech community. Through rater-perceived comprehensibility ratings, this study documents higher comprehensibility among girls than boys. It also documents higher levels of oral L2 output among transfer students, which included some from schools where English was a social and academic language, as compared to students schooled exclusively at ISSA where English was an academic language but not the students‘ social language. Conducted during naturally-occurring school events (recordings at the 12th-graders‘ speech event and ratings at the teachers‘ professional development event), our study used certain procedures from laboratory-based comprehensibility studies. Such studies had identified similarities among listeners when evaluating nonnative utterances despite factors that could have influenced their responses, e.g., linguistic properties of stimuli, L1 background of listeners, and bias against accents (Munro et al., 2006). Because these studies had shown comprehensibility to be a valid construct, the current study used comprehensibility as a means (as opposed to ―an end‖) for examining how gender and schooling might contribute to L2 variability among immersion learners in a specific setting. Hence, this study brings an expanded pedagogical application to this field of inquiry.

Looking More Deeply at Gender-Related Findings

This study found that gender significantly contributed to the oral L2 comprehensibility of ISSA immersion learners, with females being more

comprehensible than males. Care was taken not to interpret gendered comprehensibility scores as indicative of successful or unsuccessful language learning. The mean score of 5.34 (on a scale of 1 to 7, from lowest to highest comprehensibility) suggests that most of these 30 ISSA students were indeed quite easy to understand. However, we are concerned about the 7 students in the lowest quartile of comprehensibility, those who scored below 4.71. This quartile included all 5KC boys and 2 transfer boys. Based on perceived comprehensibility ratings from 12 teachers, these 7 boys were not as easy to understand.

These gender results can be explained through theories about covert prestige (Trudgill, 2000), solidarity and accommodation (Giles &Coupland, 1991), identity and investment (Norton, 1997), and social-cultural positioning (Eckert & McConnell-Ginet, 2003). We believe that some boys (especially those schooled exclusively at ISSA) tended to accommodate their speech to achieve solidarity and group distinctiveness. Such accommodations were possibly based on covert prestige that they perceived in their own nonstandard variant of English, thus making them less comprehensible to NES raters. In contrast, many girls seemed attracted to more native-like L2 speech and preferred the prestige and possible comfort of a more standard variety, making them more comprehensible to NES raters. It is uncertain whether social approval played a role for the girls.

If identity is a factor, as we strongly believe, then further questions arise: Is it within the ethical or instructional purview of the teacher to foster girls‘ or boys‘ identity formation in a particular way to influence comprehensibility? Is it even possible for the teacher to do so? In what specific ways is identity formation related to motivation? How do girls and boys interpret prestige?

Another avenue of speculation concerns the potentially differential use of learning strategies by ISSA girls and boys. Could it be that girls, contrasted with boys, used certain learning strategies leading to greater comprehensibility? Might learning strategies have been a hidden factor favoring girls‘ comprehensibility? Numerous other studies in many countries and settings have found gender differences in learning strategy use influencing proficiency, often but not always favoring females as frequent strategy users (Oxford, 1996). Further research on this topic is needed in relation to gender-related effects of learning strategy use on comprehensibility in relatively closed speech communities, such as ISSA. In the future, perhaps teachers and researchers in immersion settings should explore offering more L2 strategy instruction for all language learners, but especially for boys. However, this possibility raises additional questions, such as: What learning strategies are most commonly used by the immersion learners, often girls, who attain higher levels of oral L2 comprehensibility? Would boys in general adopt such learning strategies if presented through strategy instruction?

Schooling, L2 Proficiency, and Language Learning Success

This study found that schooling significantly contributed to ISSA students‘ oral L2 comprehensibility, with transfer students being more comprehensible than non-transfer students. We believe that the non-native oral L2 comprehensibility of many immersion ISSA learners, specifically boys, should not be viewed as unsuccessful language learning. On the contrary, our findings are consistent with U.S. and Canadian immersion studies, in which most students did not reach near-native L2 proficiency in productive skills (Swain, 1985; Swain & Johnson, 1997; Tarone& Swain, 1995). Like these prior studies, which had used other oral language assessments, our study, which used rater-perceived comprehensibility, suggests that immersion learners, if schooled exclusively at a given school, might not attain native-like oral L2 proficiency. Nonetheless, as indicated by their comprehensibility scores, the ISSA students (considered as a whole group) have been quite successful at learning the target language. Similar to students in other types of immersion studies, ISSA students acquired a somewhat high functional L2 proficiency but not necessarily near-native (Lyster 1999). Variability in oral L2 comprehensibility suggests that comprehensible input (Krashen & Terrell, 1983) is not sufficient for explaining language learning in immersion education.

Implications for Future Research

A salient contribution from this study is the way it models a practical and insightful research paradigm within a school setting. Authentic speech from a naturally-occurring academic event (i.e., student speeches) was used to determine comprehensibility. Readily-accessible devices were used for recording and playing these student speeches. Selected from a staff of experienced teachers, raters judged speech samples during a faculty development session that added to their knowledge of language learning. Based on this successful school-based adaptation of procedures usually used in laboratory-like settings, researchers and practitioners may wish to replicate this method of inquiry at other educational institutions for examining multiple issues related to oral L2 variability as well as conducting overall pronunciation assessments (Spezzini, Barratt, & Carter, 2018). Such efforts would serve to answer Derwing and Munro‘s (2005) call for greater collaboration between researchers and practitioners in conducting classroom-relevant comprehensibility studies aimed at helping all L2 learners, boys and girls, to reach higher levels of oral L2 attainment. Like the present study, these future efforts would also be a positive response to Rajadurai‘s (2007) critique of decontextualized, artificial studies conducted in laboratory-based settings.

NOTES

[1]The authors thank ISSA students and staff for their participation.

[2]We recognize the political and sociocultural difficulties surrounding labels such as NES and NNES but find these labels helpful if used carefully. Noteworthy is that, though unaware that an NES student would be on the recording, several raters smiled upon hearing her voice, and some even commented afterwards.

[3]In a larger study (Spezzini, 2004), correlations were analyzed between comprehensibility and numerous individual and social dimensions of language learning and use.

References

Abello-Contesse, C., Chandler, P. M., Lopez-Jimenez, M.D., & Chacon-Beltran, R. (Eds.). (2013). *Bilingual and multilingual education in the 21st century*. Bristol, UK: Multilingual Matters.

Aron, A., Aron, E., &Coups, E. (2005). *Statistics for the behavioral and social sciences: A brief course*, 3rd ed. Upper Saddle River, NJ: Pearson Prentice-Hall.

Baker, C., &Jones, S. (1998). *Encyclopedia of bilingualism and bilingual education.* Clevedon, UK: Multilingual Matters.

Bergeron, A., &Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals, 50*(3), 547-566.

Brantmeier, C., Schueller, J., Wilde, J., &Kinginger, C. (2007). Gender equity in foreign and second language learning. In S. S. Klein (Ed.), *Handbook for achieving gender equity through education*, (pp. 305-334). New York, NY: Routledge.

Brown, H. D. (2006). *Principles of language learning and teaching*, 5th ed. White Plains, NY: Pearson-Longman.

Carder, M. (2007). *Bilingualism in international schools.* Clevedon, UK: Multilingual Matters.

Cloud, N., Genesee, F., &Hamayan, E. (2000). *Dual language instruction: A handbook for enriched education.* Boston, MA: Heinle&Heinle.

Crowther, D., Trofimovich, P., &Issacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility? *Journal of Second Language Pronunciation, 2*(2), 160-182.

Crowther, D., Trofimovich, P., Saito, K., &Issacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background? *TESOL Quarterly, 49*, 814-837.

deCourcy, M., (2002). *Learners' experiences of immersion education.* Clevedon, UK: Multilingual Matters.

deMejía, A. (1998). Bilingual storytelling: Code switching, discourse control, and learning opportunities. *TESOL Journal*, *7*(6), 4-10.

de Mejia, A. (Ed.). (2005). *Bilingual Education in South America*. Clevedon, UK: Multilingual Matters.

Derwing, T., &Munro, M. (1997). Accent, comprehensibility and intelligibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(1), 1-16.

Derwing, T., &Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*(3), 379-397.

Derwing, T., &Munro, M. (2010, March). *L2 pronunciation and fluency development: A 7-year longitudinal study.* Paper presented at the American Association for Applied Linguistics conference, Atlanta, GA.

Derwing, T., Munro, M., &Thomson, R. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics*, *29*(3), 359-380.

Derwing, T., Rossiter, M., Munro, M., &Thomson, R. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*(4), 655–679.

Derwing, T. M., Thomson, R.I, & Munro, M. (2006). English pronunciation and fluency development in Mandarin and Slavic speakers. *System, 34,* 183-193.

Eckert, P., &McConnell-Ginet, S. (2003). *Language and gender*. Cambridge, UK: Cambridge University Press.

Ellis, R. (1994). *The study of second language acquisition.* Oxford, UK: Oxford University Press.

Gass, S., Sorace, A., &Selinker, L. (1999). *Second language learning data analysis*, 2nd ed. Mahwah, NJ: Erlbaum.

Giles, H., &Coupland, N. (1991). *Language: Contexts and consequences*. Keynes, UK: Open University Press.

Green, J., & Oxford, R. (1995). A closer look at learning strategies, L2 proficiency, and gender. *TESOL Quarterly*, *29*(2), 261-297.

Gumperz, J. (1992). Contextualization and understanding. In A. Duranti& C. Goodwin (Eds.), *Rethinking context* (pp. 229–252). Cambridge, UK: Cambridge University Press.

Hatch, E., &Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York, NY: Newbury House.

Hayden, M., Levy, J., & Thompson, J. (2015). *The SAGE handbook of research in international education* (2nded.). Los Angeles, CA: SAGE.

Hayes-Harb, R., & Hacking, J. (2015). What do listeners believe underlies their accentedness judgments? *Journal of Second Language Pronunciation, 1* (1), 43-64.

Issacs, T., &Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*, 475-505.

Johnson, R. K., & Swain, M. (1997). *Immersion education: International perspectives.* New York, NY: Cambridge University Press.

Jun, H. G., & Li, J. (2010). Factors in raters' perceptions of comprehensibility and accentedness. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference,* Iowa State University. Sept. 2009. (pp. 53-66). Ames, IA: Iowa State University.

Kachru, B., & Nelson, C. (1996). World Englishes. In S. McKay & N. Hornberger (Eds.), *Sociolinguistics and language teaching,* (pp. 71-102). Cambridge, UK: Cambridge University Press.

Kennedy, S., &Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review, 64*(3), 459-489.

Krashen, S., &Terrell, T. (1983). *The natural approach: Language acquisition in the classroom.* Hayward, CA: Alemany Press.

Labov, W. (1970). *The study of nonstandard English.* Urbana, IL: National Council of Teachers of English.

Levis, J. M. (2018). Precision and imprecision in second language pronunciation. *Journal of Second Language Pronunciation, 4*(1), 1-10.

Lyster, R. (1999). Immersion. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics*, (pp. 626-632). Amsterdam, NL: Elsevier.

Major, R. (2004). Gender and stylistic variation in second language phonology. *Language Variation and Change, 16*(3), 169-188.

Major, R., Fitzmaurice, S., Bunta. F., &Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessments. *TESOL Quarterly, 36*(2), 173-190.

Motallebzadeh, K., &Nematizadeh, S. (2011). Does gender play a role in the assessment of oral proficiency? *English Language Teaching, 4*(4), 165-172.

Mougeon, R., & Rehner, K. (2001). Acquisition of sociolinguistic variants by French immersion students: The case of restrictive expressions, and more. *Modern Language Journal, 85*(3), 398-415.

Moyer, A. (2010, March). *Do gender differences in L2 accent really exist?* Paper presented at the American Association of Applied Linguistics conference, Atlanta, GA.

Munro, M., &Derwing, T. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning,*45(1), 73-97.

Munro, M., &Derwing, T. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition, 23*(4), 451-468.

Munro, M., &Derwing, T. (2015). A prospectus for pronunciation research in the 21st century. *Journal of Second Language Pronunciation, 1*(1), 11-42.

Munro, M., Derwing, T., &Morton, S. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*(1), 111-131.

Munro, M., & Mann, V. (2005). Age of immersion as predictor of foreign accent. *Applied Psycholinguistics, 26,* 311-341.

Norton, B. (1997). Language, identity, and the ownership of English. *TESOL Quarterly 31*(3), 409-430.

Norton, B., &Pavlenko, A. (2004). Addressing gender in the ESL/EFL classroom. *TESOL Quarterly 38*(3), 504-513.

O'Brien, M.G. (2014). L2 learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning, 11,* 715-748.

Oxford, R. L. (1995). Gender differences in language learning styles: What do they mean? In J. M. Reid (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 34-46). Boston, MA: Heinle&Heinle.

Oxford, R. L. (Ed.). (1996). *Language learning strategies around the world: Crosscultural perspectives.* Manoa, HI: University of Hawaii Press.

Parlak, O. (2010, March). *Does pronunciation instruction promote intelligibility and comprehensibility?* Paper presented at the Teachers of English to Speakers of Other Languages conference, Boston, MA.

Pavlenko, A. (2008). Research methods in the study of gender in second/foreign language education. In K. A. King & N. H. Hornberger (Eds.), *Encyclopedia of languageand education: Vol. 10. Research methods in language education* (2nd ed., pp. 165-174). New York, NY: Spring Science+Business Media LLC.

Pennington, M. (1996). *Phonology in English language teaching.* London, UK: Longman.

Pickering, L. (2006). Current research on intelligibility in English as a lingua franca. *Annual Review of Applied Linguistics 26,* 219-233.

Potowski, K. (2007). *Language and identity in a dual immersion school*. Clevedon, UK: Multilingual Matters.

Rajadurai, J. (2007). Intelligibility studies: A consideration of empirical and ideological issues. *World Englishes 26*(1), 87-98.

Saito, K., &Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly, 11,* 421-446.

Saito, K., Trofimovich, P., &Issacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*(11),217-240.

Schneider, F. W., Gruman, J. A., & Coutts, L. M. (Eds.). (2005). *Applied social psychology: Understanding and addressing social and practical problems.* Thousand Oaks, CA: Sage.

Smith, L., &Bisazza, J. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning 32*(2), 259-269.

Southwood, M. H., &Flege, J. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Clinical Linguistics and Phonetics 13*(5), 335-349.

Spezzini, S. (2004). English immersion in Paraguay: Individual and sociocultural dimensions of language learning and use. *International Journal of Bilingual Education and Bilingualism, 7*(5), 412-431.

Spezzini, S., Barratt, L., & Carter, D. (2018). Pronunciation assessment. In C. Coombe (Ed.), *Assessment and evaluation, Volume 8* (pp. 5174-5180), in J. I. Liontas (Editor-in-Chief), *The TESOL Encyclopedia of English Language Teaching*. Hoboken, NJ: John Wiley & Sons, Inc.

Spezzini, S., & Oxford, R. (2003, March). *Immersion L2 usage: Issues of identity, gender, and ethnicity.* Paper presented at the American Association for Applied Linguistics conference. Arlington, VA.

Strachan, L., Kennedy, S., &Trofimovich, P. (2019, March 14). Second language speakers‘ awareness of their own comprehensibility. *Journal of Second Language Pronunciation.* John Benjamins Publishing Company. E-issn 2215-194x.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass&
C.Maden (Eds.), *Input in second language acquisition* (pp. 235-256). Rowley, MA: Newbury House.

Swain, M., &Johnson, R. (1997). Immersion education: A category within bilingual education. In R. Johnson & M. Swain (Eds.), *Immersion education: International perspectives*, 1-16. New York, NY: Cambridge University Press.

Swann, J. (1999). Gender and language. In B. Spolsky (Ed.), *Concise encyclopedia of educational linguistics* (pp. 200-202). Amsterdam, NL: Elsevier.

Tannen, D. (1990). *You just don’t understand: Men and women in conversation.* New York, NY: Morrow.

Tarone, E., &Swain, M. (1995). A sociolinguistic perspective on second language use in immersion classrooms. *Modern Language Journal, 79*(2), 166-178.

Trudgill, P. (2000). *Sociolinguistics: An introduction to language and society,* 4th ed. London, UK: Penguin Books.

van der Slik, F.W.P., van Hout, R. W. N. M., &Schepens, J. J. (2015). The gender gap in second language acquisition: Gender differences in the acquisition of Dutch among immigrants from 88 countries with 49 mother tongues. *PLoS ONE, 10*(11), 1-22.

Varonis, E., &Gass, S. (1982). The comprehensibility of nonnative speech. *Studies in Second Language Acquisition, 4*(2), 114-136.

Winke, P., &Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762-789.

Zhu, X. (2009). *Gender dynamics in peer interaction and their influence on second language learning in the English-as-a-second language classroom.* (Unpublished doctoral dissertation). University of South Florida, Tampa, FL.