Studies in Second Language Acquisition, 2014, **37**, 1–28.
doi:10.1017/S0272263114000576

AGE MATTERS, AND SO MAY RATERS

Rater Differences in the Assessment of Foreign Accents

QA
Becky H. Huang
University of Texas at San Antonio

Sun-Ah Jun
University of California Los Angeles

Research on the age of learning effect on second language learners' foreign accents utilizes human judgments to determine speech production outcomes. Inferences drawn from analyses of these ratings are then used to inform theories. The present study focuses on rater differences in the age of learning effect research. Three groups of raters who differed in their native language background and language experience participated in the study: inexperienced native English speaker (NES) raters, experienced NES raters, and advanced nonnative English speaker (NNES) raters. All raters evaluated 64 speech samples taken from both NESs and NNESs who varied in their age of arrival in the second-language-speaking country. Results from the study showed that experienced NES raters were better than the other two rater groups at distinguishing NESs from NNESs. Although no rater group differences were found in scaling speakers' foreign accents or in

This study was supported by a postdoctoral research grant from the National Science Council of Taiwan and a faculty seed grant from the University of Texas at San Antonio to the first author, and a UCLA Faculty Research Grant to the second author. We thank the four anonymous *SSLA* reviewers and Jason Bishop for their constructive and detailed comments on the drafts of the study. We are also grateful to all the participants in this study for their time and to Daniel Sass for his assistance with data analysis. All remaining errors are our own.

Correspondence concerning this article should be addressed to Becky Huang, Department of Bicultural-Bilingual Studies, The University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249. E-mail: becky.huang@utsa.edu

their interrater reliability, inexperienced NES raters were stricter in their ratings than both experienced NES raters and advanced NNES raters. Implications for research on age of learning effects are discussed.

Research on the effect of age on learning a second language (L2) has consistently shown that L2 learners who are immersed in a L2 at a younger age produce more targetlike phonology—in particular, more nativelike accents—than those immersed in the L2 later in their life (Asher & García, 1969; Flege, Munro, & MacKay, 1995; Flege, Yeni-Komshian, & Liu, 1999; Huang & Jun, 2011; Moyer, 1999, 2004; Munro & Mann, 2005; Oyama, 1976). The age of learning variable remains the most powerful predictor of a nativelike accent, even after controlling for potential confounding variables, such as L2 learners' educational attainment and their amount of exposure to the L2 (Flege et al., 1999). Researchers have argued that L2 phonological production is strongly constrained by the age of learning variable because of the neuromuscular demands involved in producing targetlike phonology (Scovel, 2000). To produce nativelike phonology, it is hypothesized that the cutoff age is around 6 for segments and 12 for suprasegments (Long, 2005).

Empirical studies that examine the age effect on L2 phonological production generally use human judgments to determine L2 learners' outcomes, particularly in the domain of the degree to which learners have foreign accents (FAs). On the basis of the inferences drawn from analyses of raters' ratings, researchers then test or formulate theories of language learning. However, although research in the fields of speech processing and oral language assessment has revealed robust differences among raters, rater differences have received little attention in research on the age effect. In the following sections we first summarize the speech processing and language assessment literature before reviewing in more detail studies relating specifically to the focus of the present study—namely, age of learning effects and the effects of rater differences. We then describe the methodology of the current study and report the results. We conclude with a discussion of the implications of rater differences for the age of learning research and provide suggestions for future research.

BACKGROUND

Rater Differences in Speech Processing and Oral Language Assessment

Research on speech processing and oral language assessment has generally shown that raters' or listeners' native language background and

language experience can influence their perception of FAs, their ratings of the intelligibility and comprehensibility of nonnative speech, and their judgments of nonnative speakers' (NNSs') language proficiency (e.g., Bradlow & Bent, 2003; Brodkey, 1972; Gass & Varonis, 1984; Huang, 2014; Winke & Gass, 2013; Winke, Gass, & Myford, 2013; Xi & Mollaun, 2011). The influence of rater[1] differences has been an important topic in the study of speech processing because this topic addresses fundamental questions about human perception and learning (see Carey, Mannell, & Dunn, 2011, for a thorough review of dominant theories). It is also one of the most heavily surveyed areas in oral language assessment due to its direct implications for the reliability, validity, and fairness of assessments (Bachman, Lynch, & Mason, 1995; Kunnan, 2000).

Turning first to the effects of raters' native language background (native speaker [NS] vs. NNS) on ratings of nonnative speech, results from empirical research are quite mixed. In terms of rank ordering nonnative speech, Derwing and Munro (2013), for example, found that the rank-ordering patterns of highly proficient NNS raters closely paralleled those of NS raters. Ratings by highly proficient NNS raters also significantly correlated with those of NS raters. In terms of the numerical ratings assigned to nonnative speech, some studies have found NS raters to be more lenient than their NNS counterparts (e.g., Caban, 2003; Fayer & Krasinski, 1987), whereas others have observed no significant differences between the two groups (Kim, 2009; Zhang & Elder, 2011). However, analyses of interviews and raters' written comments have revealed that NS and NNS raters may weigh various features of speech samples differently in the justifications of their ratings (Kim, 2009; Zhang & Elder, 2011). To illustrate, Kim (2009) conducted a mixed-methods study to assess how NS and NNS teachers evaluate students' oral English performance. Both NS and NNS teachers judged 10 Korean students' English speech samples on a 4-point scale and provided written comments to justify their ratings. Although quantitative analyses of teacher ratings revealed no significant group differences, NS teacher raters commented most frequently on overall language use, pronunciation, and vocabulary, whereas NNS teachers emphasized pronunciation, vocabulary, and intelligibility. Native speaker raters also provided more elaborate comments on speakers' pronunciation and grammar than their NNS peers, suggesting potential qualitative differences in the two groups' rating behavior.

An additional problem in investigating the influence of rater experience concerns how experience is defined and operationalized, which varies substantially across studies (Isaacs & Thomson, 2013). Some studies have defined it as the exposure to foreign-accented speech in controlled experiments (Bradlow & Bent, 2003, 2008; Clarke & Garrett, 2004), some have appealed to raters' professions (e.g., second or foreign language teachers or speech therapists vs. undergraduate students

without any linguistics training) as a proxy (Carey et al., 2011; Kennedy & Trofimovich, 2008), some have relied on participants' self-reports of familiarity with foreign accents (Munro, Derwing, & Morton, 2006; Winke et al., 2013), and still others have used some combination of these different criteria (Huang, 2014). There is also much variation in basic research design across studies. Whereas some have utilized experimental methodology to test the effects of perceptual training on the intelligibility and comprehension of foreign-accented speech (e.g., Bradlow & Bent, 2003, 2008; Clarke & Garrett, 2004), others have used quasi-experimental designs to examine the effects of language experience by studying different rater groups (e.g., Brodkey, 1972; Huang, 2014; Isaacs & Thomson, 2013; Kennedy & Trofimovich, 2008; Major, 2007; Munro et al., 2006; Xi & Mollaun, 2011).

Despite differences in design and in the definition and measurement of language experience, the overall picture emerging from this body of work suggests that raters' language experience facilitates their comprehension of nonnative speech. Language experience has also contributed to modifying raters' perceptions of FAs and judgments of L2 speakers' proficiency, though the relationship between the degree of severity in the rating and the raters' experience remains unclear. Some studies have found that experienced raters tend to be more lenient than linguistically naïve raters in their ratings (Barnwell, 1989; Carey et al., 2011; Hsieh, 2011; Winke et al., 2013), but others have found the opposite to be true (Galloway, 1980; Hadden, 1991). Furthermore, studies comparing specific features used by raters (e.g., global proficiency, pronunciation, or fluency) have also shown that language experience may play a significant role in their decision-making processes (Chalhoub-Deville, 1995; Hsieh, 2011; Winke & Gass, 2013; Xi & Mollaun, 2011).

Rater Differences in the Age of Learning Effect Research AQ1

We found a total of nine studies that examined rater differences in the assessment of the age effect,[2] and all of them included rater differences as a secondary research question rather than as the primary focus.[3] There was also much variation across studies in the specific background variables examined, which included the investigation of NS raters with mainstream versus regional accents (Bongaerts, 1999b; Bongaerts, Planken, & Schils, 1995), NS raters with and without experience in teaching and linguistic training (Bongaerts, Mennen, & van der Slik, 2000; Bongaerts, van Summeren, Planken, & Schils, 1997; Nikolov, 2000, Study 1; Thompson, 1991), NNSs with limited versus substantial exposure to the target language (Flege, 1988; MacKay, Flege, & Imai, 2006), and children versus adults (Nikolov, 2000).

Bongaerts and colleagues (Bongaerts, 1999b; Bongaerts et al., 2000; Bongaerts et al., 1995; Bongaerts et al., 1997) conducted a series of empirical studies to identify highly advanced late L2 learners with nativelike proficiency. The main purpose of their work, a test of the critical period hypothesis, was to determine whether late L2 learners can "beat the odds" of a late start and achieve nativelike proficiency in their L2. In an initial study, Bongaerts et al. (1995) found that NS raters from northern England who were unfamiliar with regional British accents judged the speech of NSs of British English with regional accents to be less nativelike than that of Dutch learners of British English with training in Received Pronunciation. These curious results led the authors to conduct a subsequent study, published in the same paper, in which they substituted NSs of regional British accents with NSs of neutral accents and included NS raters both with and without experience in language teaching and linguistics. Bongaerts and colleagues compared ratings by two NS rater groups (experienced vs. inexperienced) and found that they were not reliably different from each other. Using a widely adopted *nativelikeness criterion* (Abrahamsson & Hyltenstam, 2009)—namely, the criterion of two standard deviations within the mean of the NS control group—the authors identified five late Dutch learners of British English who were perceived as speaking with a nativelike accent. With evidence of late learners having achieved nativelike proficiency, the authors argued against the critical period hypothesis. In another study, Bongaerts (1999b) examined a different first language (L1)–L2 pairing (Dutch learners of French) and again included both experienced and inexperienced NSs of French as raters. The researchers replicated their previous finding, showing no reliable differences between the judgments of the experienced and inexperienced raters. The results also revealed three adult learners who demonstrated a nativelike accent in their L2 French.

In contrast to these findings, Bongaerts and colleagues found discrepancies between the judgments of raters with and without linguistic experience in a fourth study (Bongaerts et al., 2000). In this study, two groups of Dutch NSs, experienced and inexperienced, rated the speech production of L2 learners of Dutch and of Dutch NS control participants. Experienced raters were, again, trained L2 teachers, whereas inexperienced raters lacked any experience in language teaching or linguistics. Using the nativelikeness criterion derived from experienced raters' average ratings, the researchers identified four late L2 learners with nativelike accents, but only two of them were also perceived as nativelike based on inexperienced raters' ratings. The researchers combined the results and made inferences from the two, instead of four, nativelike cases as evidence against the critical period hypothesis.

In an attempt to replicate the work by Bongaerts and colleagues, Nikolov (2000) also found raters' behavior to vary as a function of

linguistic experience. She included L2 learners of Hungarian (Study 1) and English (Study 2) as well as NSs of each target language as control participants. For both studies, Nikolov included three groups of NS raters in the respective L2, one adolescent and two adult groups. In Study 1, the adults were either university students or language teachers, whereas both groups of adults were postgraduate students in Study 2. Results from the two studies showed that adult raters rank ordered the speakers in a similar manner. Adolescents' ratings, in contrast, were not as highly correlated with the adults' ratings. Crucially, in Study 1, Nikolov found that teacher raters, compared with university students, were better able to distinguish NSs and NNSs.

In line with these studies showing positive rater differences, Thompson (1991) investigated factors predicting Russian immigrants' degrees of FA in English by assessing the judgments of two groups of English NSs varying in their experience in language teaching and their contact with Russian immigrants in the United States (experienced vs. inexperienced). The results revealed that, compared to inexperienced NS raters, experienced NS raters were more lenient and more reliable in their assessment of FAs. It is worth noting that Thompson did not use the commonly adopted nativelikeness criterion to identify nativelike L2 learners. Instead, she examined mean ratings and found two nativelike speaker participants with perfect scores.

In a study examining the effects of the amount of L1 use on Italian immigrants' L2 pronunciation outcomes, Flege, Frieda, and Nozawa (1997) also included two groups of English NSs as raters. One rater group was from Alabama, United States, and self-reported having limited exposure to FAs; the other rater group was from Ottawa, Canada, and reported having regular exposure to a variety of FAs, including Italian-accented English. To our knowledge, this is the only study in the age effect literature that includes detailed background information about the raters employed. The speaker participants in the study included two groups of Italian L1 immigrants who differed in the amount of their L1 use (high vs. low), and one control group of Canadian English NSs. Flege and his colleagues found that the two groups of raters were not reliably different in rank ordering the talkers (NS > low L1 use > high L1 use). However, raters from Ottawa were better at detecting the FAs of Italian immigrants to Canada and at distinguishing NSs and NNSs. Flege and his colleagues attributed this advantage to the Canadian raters' experience and familiarity with the specific type of FAs under investigation.

All seven of the studies just described used NS raters. The only two studies that included NNS raters were those reported by Flege (1988) and MacKay et al. (2006), who examined NNS raters' ability to gauge FAs in their L2. Flege (1988) included four speaker groups of Chinese learners of English who varied in their age of arrival (AoA) and length of residence in the United States and a group of English NSs as controls.

Three groups of raters were recruited to judge the FAs of these speakers: one English NS group and two NNS groups with 1 and 5 years of residence in the United States, respectively. Flege found that the two groups of NNS raters rank ordered the speakers in a similar manner as the NS raters. However, NNS raters with longer residence in the United States were better at distinguishing between NSs and NNSs than were NNS raters with shorter residence. The ratings of NNS raters with longer residence also did not reliably differ from those of the NS raters. The combined results suggested that experience (from longer residence) may afford the NNS raters perceptions of the L2 sounds that are more similar to those of NS raters. A major limitation of the study was that most of the raters were drawn from the speaker participants and thus rated their own speech, which raised concerns about the validity of their ratings.

Similar to Flege's (1998) study, MacKay et al. (2006) evaluated NNSs' ability to scale L2 speakers' FAs.[4] However, unlike the NNS raters in Flege's study, who shared the L1 (Chinese) of the NNS speakers, the NNS raters in MacKay et al.'s study spoke a different L1 (Arabic) than the NNS speakers (Italian). The experiment involved English NS control speakers and four groups of Italian learners of English, differing in their AoAs and L1 usage (early AoA–low L1 use; early AoA–high L1 use; late AoA–low L1 use; and late AoA–high L1 use). Two groups of raters, one English NS group and one NNS group with an Arabic L1 background, evaluated all speakers' speech samples. The NNS raters themselves had AoAs of 15 and older, and their average length of residence was 3 years. Results from the study revealed that NNS raters' ratings were highly correlated with those of NS raters, suggesting that even NNS raters who do not share the L1 of the NNS speakers can reliably scale the speakers' FAs.

In summary, existing studies paint a mixed picture about the role of language experience in rating behavior;[5] some have suggested that NS raters' language experience may lead them to be more lenient (Bongaerts et al., 2000; Nikolov, 2000; Thompson, 1991) or may help them distinguish NSs from NNSs (Flege et al., 1997; Nikolov, 2000). Other studies, however, have found no such effect (Bongaerts 1999b; Bongaerts et al., 1995; Bongaerts et al., 1997). Two studies that compare NNS raters with NS raters have found that NNS raters can scale the FAs of L2 speakers as well as their NS peers (Flege, 1988; MacKay et al., 2006), but these findings await replication. It is also unknown how NNS raters compare to NS raters in terms of the degree of severity in their judgments and their ability to distinguish between NSs and NNSs.

By systematically investigating the effect of rater differences on the assessment of FAs, the present study aims to clarify how raters' language experiences affect their judgments of foreign accents. In particular, we focus on raters' L1 background and their experience with FAs and language teaching, using a quasi-experimental design. This included comparison of three groups of raters: advanced nonnative English

speakers (advanced NNESs), inexperienced native English speakers (inexperienced NESs), and experienced native English speakers (experienced NESs). Our research questions focused on whether raters' L1 and linguistic experience influenced any of the following:

1. The reliability of their foreign accent ratings.
2. Their scaling of L2 speakers' degree of foreign accents.
3. The degree of severity in their ratings of L2 speakers' FAs.
4. Their ability to reliably distinguish NSs from NNSs.

METHOD

Participants: Speakers

Four groups of speakers participated in the study: three L2 learner groups, varying in AoA (child arrivals, adolescent arrivals, and adult arrivals), and one NES control group (n = 14). The L2 learner groups immigrated to the United States either in childhood (ages 5–11; n = 22), adolescence (ages 12–16; n = 14), or adulthood (ages 17–25; n = 14) and had lived in the United States for at least 7 years. They all spoke Mandarin Chinese as their L1 and were at least college educated or were undergraduate students at the time of testing. To control for amount of L2 input, we matched the three groups on their length of residence (LoR) in the United States and their self-reported English input in the past 5 years (estimated as a percentage) in three different domains (oral, media, and literacy).

All of the NES control speaker participants were NSs of American English and had no significant exposure to a L2 other than formal foreign language instruction in high school. To avoid any dialectal differences, we screened and selected participants without a noticeable regional accent. See Table 1 for a summary of the demographic information for all speaker participants.

Participants: Raters

Three groups of raters participated in the current study: advanced NNES, inexperienced NES, and experienced NES raters. All raters were at least college educated or were current undergraduate students at the time of testing. They reported normal hearing and no history of hearing or speech problems.

We carefully screened all potential rater participants via a survey instrument before inviting them to participate in the study. We used

Table 1. Means and Standard Deviations (in Parentheses) of Demographic Information by Speaker Group

Demographic variable	NES controls (n = 14)	L2 child arrivals (n = 22)	L2 adolescent arrivals (n = 14)	L2 adult arrivals (n = 14)
Chronological age	27 (7)	24 (5)	29 (6)	36 (5)
Age of arrival	NA	8.55 (1.77)	13.93 (1.64)	21.79 (2.83)
Gender	7f, 7m	15f, 7m	4f, 10m	11f, 3m
Length of residence	NA	15.18 (4.65)	14.74 (6.04)	13.58 (4.87)
English input-oral	NA	71 (15)	75 (14)	68 (14)
English input-media	NA	81 (17)	83 (20)	74 (24)
English input-literacy	NA	90 (10)	87 (10)	81 (23)

Note. NES = native English speaker; f = female; m = male; NA = not applicable. Chronological age = current age in years; age of arrival = age at which the participant arrived in the United States; length of residence = length of stay in the United States in years; English input-oral = self-estimated English exposure in oral domain (i.e., listening or speaking) in the past 5 years in percent form; English input-media = self-estimated English exposure via media (i.e., TV, radio, movies) in the past 5 years in percent form; English input-literacy = self-estimated English exposure in literacy domain (i.e., reading or writing) in the past 5 years in percent form. One-way ANOVAs revealed no significant differences among the three groups in LoR, $F(2, 47) = 0.42$, $p = .657$, and English input in the oral domain, $F(2, 47) = 2.98$, $p = .060$, the media domain, $F(2, 47) = 1.09$, $p = .346$, or in literacy domain, $F(2, 47) = 1.88$, $p = .165$.

different screening criteria for each of the three groups. Advanced NNESs were required to (1) speak Mandarin Chinese as their L1, (b) have come to the United States after college and have lived in the United States for at least 7 years, (c) self-report high levels of oral English proficiency (at least a rating of 7 on a scale of 1–9), and (d) self-report high levels of familiarity with FAs (at least a rating of 3 on a scale of 1–4). Almost half of the advanced NNES raters had also taught English as a foreign language in their home countries. Both inexperienced NES and experienced NES rater groups were required to be NSs of American English and to speak without a noticeable regional accent and to have had no significant exposure to a L2 prior to high school. Inexperienced NES raters also met the following criteria: (a) They did not speak any L2 or foreign language fluently at the time of participation, (b) they had not taken any class related to speech or phonetics, (c) they lacked any experience in language teaching or the rating of nonnative speech, and (4) they self-reported little familiarity with FAs. To qualify for the study, experienced NES raters were required to (a) be current or past English as a second language (ESL) teachers or speech therapists,[6] (b) have worked in that profession for at least one full year, and (c) self-report high levels of familiarity with FAs (at least a rating of 3 on a scale of 1–4). Three of the experienced NES raters also had knowledge of Mandarin Chinese via either formal language classes or self-study. See Table 2 for a summary of the demographic information of all rater participants.

Table 2. Means and Standard Deviations (in Parentheses) of Demographic Information by Rater Group[a]

Demographic variable	Advanced NNES (n = 10)	Inexperienced NES (n = 10)	Experienced NES (n = 10)
Chronological age	32 (3)	24 (3)	31 (4)
Gender	5f, 5m	5f, 5m	5f, 5m
Length of residence	9 (3)	NA	NA
English oral proficiency	7.10 (0.74)	NA	NA
Familiarity with foreign accents	3.10 (0.32)	2 (0)	3.50 (0.53)
TESL/TEFL experience	0.50 (0.85)	0	80.80 (1.95)

Note. NNES = nonnative English speaker; NES = native English speaker. English oral proficiency = self-rated English oral proficiency on a 1-9 scale (*1* = poor, *9* = nativelike); familiarity with foreign accents = self-rated familiarity with foreign accents on a 1-4 scale (*1* = not familiar at all. I am not able to tell whether the speaker's first language is English or other languages, *2* = somewhat familiar. I can sometimes tell whether the speaker's first language is English or other languages, *3* = moderately familiar. I can often tell whether the speaker's first language is English or other languages, *4* = very familiar. I can always tell whether the speaker's first language is English or other languages); TESL/TEFL experience = years of experience in teaching learners of English as a second/foreign language. One-way ANOVA revealed a significant group effect, $F(2, 27) = 47.90$; $p < .001$, and Bonferroni post hoc tests suggested that both advanced NNES and experienced NES groups reported significantly higher familiarity with foreign accents than inexperienced NES group (both $p < .001$). Experienced NES raters reported slightly higher familiarity than advanced NNES raters, though the difference was only marginally significant ($p = .054$).

Procedure

Collection of Speech Samples. All speakers were recorded reading a paragraph in a quiet room at a university laboratory or their private residences (Weinberger, 2014; see the Appendix for the paragraph). They were given 1 min to review the paragraph before recording and were instructed to read at their own pace. Recordings were completed using a high-quality head-mounted microphone (Shure SM 10A) in the program Audacity (Audacity Team, 2000, Version 1.2.5) on a laptop (IBM Thinkpad x60s).

Collection of Ratings. We normalized the intensity of all speech files (75 dB) using PRAAT (Version 5.0, Boersma & Weenink, 2009) and assigned an identification number by the order of speakers' AoAs. We then generated two lists of randomized orders using an online number randomizer (Urbaniak & Plous, 2013). The two orders were then balanced within each rater group. The average length of the speech files was approximately 28 s (range = 22–40).

Each rater participant rated all 64 speech files in a single 30–40 min session. There were two data-collection sites: one was a university laboratory and the other was a quiet office in a research institution. Following previous research (Bongaerts, 1999b; Bongaerts et al., 1997),

rater participants were not informed of the purpose of the study but were told that they would listen to and rate speech samples produced by unspecified proportions of NSs and NNSs of American English. No practice or training was provided. Rater participants listened to each speech file via headphones (Sennheiser HD 203) on a laptop (IBM Thinkpad x60s). Speech files were played via the MATLAB program (Version 6.5.1). Rater participants were instructed to judge the degree of FAs using a 1–9 scale (*1* = strong FA, *9* = native English speaker).

To standardize the rating procedure and condition across the two sites, we used the same models of headphones and laptops and the same presentation program. The two researchers working on the project also adhered to the same script when giving instructions to the rater participants. The distribution of raters from the three rater groups was roughly equal between the two sites.

The current study presents several methodological improvements from previous research. First, to understand the comprehensive picture of rater effects, the study included speakers with a range of AoAs; second, it also controlled for speakers' input, which was operationalized as a combination of LoR and amount of L2 input. Previous studies either focused only on adult learners (Bongaerts, 1999b; Bongaerts et al., 2000; Bongaerts et al., 1995; Bongaerts et al., 1997; Nikolov, 2000) or included child and adult learners, but not adolescent learners (Flege, 1988; MacKay et al., 2006). In an earlier study that covered a range of AoAs (Thompson, 1991), the researcher did not control for learners' input, making the interpretations of the results difficult. The current design permits investigations of the potential interactions between speakers' AoAs and differences in behavior among raters. Additionally, we carefully screened the raters to ensure they met our selection criteria, and we used a larger sample of raters than that of most previous studies. We also chose a paragraph that contained the full inventory of American English phones to elicit read-aloud production. Compared to the sentence-level productions used in previous research, the production of paragraph-length samples provided more speech materials for raters to evaluate, which in turn afforded us better chances to identify possible differences among raters' perceptions. Finally, the current study adopted a broader definition of language experience and used a combination of both objective (raters' profession) and subjective (self-reports of familiarity with FAs) measures of language experience.

RESULTS

To examine interrater reliability and determine whether it was statistically appropriate to combine rater participants' FA ratings, we calculated average absolute agreement using a two-way mixed intraclass

correlation (ICC) for each rater group. The ICC measures were .92 (95% CI [.88, .95]) for the advanced NNES raters, .95 (95% CI [.92, .97]) for the experienced NES raters, and .97 (95% CI [.95, .98]) for the inexperienced NES raters. The ICCs were all very high across rater groups and were comparable to the values reported in previous studies that used ICCs as an interrater reliability index (e.g., Derwing & Munro, 2013; MacKay et al., 2006; Munro et al., 2006; Trofimovich, Lightbown, Halter, & Song, 2009). We thus computed the average of these ratings within each group to use as dependent variables in the statistical models. The confidence intervals suggest a narrow interval around the ICC point estimates and indicate that the only difference (based on the confidence intervals) was between advanced NNES and inexperienced NES raters.[7] We also examined the correlation matrix to substantiate the ICCs, and we found that the coefficients were consistent and large (smallest r = .49) within each rating group. However, the average Spearman's rho correlation coefficients were smaller for the advanced NNES raters (ρ_{mean} = .61, s = .01) than for the experienced NES raters (ρ_{mean} = .86, s = .01) and the inexperienced NES raters (ρ_{mean} = .87, s = .01).

We then conducted the traditional 4 (between, or speaker group, factor) × 3 (within, or rater group, factor) MANOVA to assess the effect of raters on the dependent variable, rater FA ratings. This multivariate model rests on the following assumptions: homogeneity of variance, equality of covariance matrices, multivariate normality, and independent errors. Although preliminary analyses revealed these assumptions were met when excluding the NES speaker group from the analyses, the inclusion of this group violated the equality of covariance matrices and the homogeneity of variance assumptions, as the estimated standard deviations were much smaller for this group. These findings were expected, given that the NES speaker group should have much higher ratings and less variability among raters than the other speaker groups. Consequently, including this group in the model biases the test statistics (i.e., it underestimated the standard errors and, therefore, the test statistics) and statistical inferences by violating the statistical assumptions of the model (i.e., overall MANOVA and pairwise comparisons). Although one conventional approach is to transform the dependent variable, these assumptions remained violated after conducting Box-Cox transformations. To resolve these concerns, we elected to employ robust analyses that use nonpooled standard errors and did not rely on traditional model assumptions.

Empirical evidence suggests that the approximate degrees of freedom (ADF) tests are relatively robust for tests of mean equality when the aforementioned assumptions are violated (for a review, see Keselman, Algina, & Kowalchuk, 2001; Lix & Keselman, 1995). Moreover, the statistical power of these ADF tests is generally comparable to traditional multivariate and univariate mixed ANOVA approaches, even when model assumptions are satisfied. Especially important for this study is

the observation that the ADF tests are also more powerful than traditional tests when both variances and group cell sizes are unequal. For these reasons, we conducted a 4 (speaker group) × 3 (rater experience) mixed ADF ANOVA followed by robust Welch-James ADF pairwise mean comparisons. Recall that the Welch-James ADF pairwise comparison procedure does not pool the variances; thus the smaller variances for the NES speakers will not influence the standard errors when comparing means across the other groups.

Effect sizes (i.e., Cohen's *d*) were calculated differently depending on whether the pairwise comparison was a between- or within-groups comparison, as within-groups comparisons should take into account the correlation between variables (Lipsey & Wilson, 2001). Therefore, the between-group effect sizes were calculated by taking the mean difference ($M_1 - M_2$) and dividing by the pooled standard deviation (s_p). The within-group effect sizes were computed using the following equation: $t_c[2(1 - r)/n]^{1/2}$, where t_c is the *t* statistic (or square root of the *F* statistic in Tables 3 and 4), *r* is the correlation coefficient, and *n* is the sample size (see Dunlap, Cortina, Vaslow, & Burke, 1996). AQ2

Approximate degrees of freedom mixed ANOVA analyses revealed a statistically significant main effect of rater group, $F(2, 38.63) = 225.19$, $p < .001$, $\eta_p^2 = .87$, and speaker group, $F(3, 116.20) = 116.20$, $p < .001$, $\eta_p^2 = .69$, and a rater experience by speaker group interaction, $F(6, 32.45) = 45.09$, $p < .001$, $\eta_p^2 = .40$. Based on Cohen's (1988) tentative standards (small $\eta_p^2 = .01$; medium $\eta_p^2 = .06$, large $\eta_p^2 = .14$), the effect size was large for each of the main effects and interaction. To better understand these differences, in particular the significant interaction, Welch-James ADF pairwise comparisons using a Bonferroni adjustment ($\alpha = .05/21 = .002$) were conducted. Note that, to reduce the probability of a Type I error, not all mean comparisons were conducted. Instead, only means of interest were tested for statistical significance (see Tables 3 and 4).

Pairwise analyses comparing differences among speaker groups at each rater group (see Figure 1 and Table 3) revealed that NES speakers received higher ratings than child arrivals, and child arrivals received higher ratings than adolescent arrivals, regardless of rater groups. However, no difference between adolescent arrivals and adult arrivals emerged, regardless of rater background. When comparing rater groups at each speaker group level (see Figure 1 and Table 4), analyses indicated that advanced NNES and experienced NES raters overall assigned higher scores than inexperienced NES raters, with a much smaller difference between advanced NNES and experienced NES raters. It is worth noting that experienced NES raters assigned significantly higher ratings to NESs than both inexperienced NES and advanced NNES raters, but the ratings for NESs did not significantly differ between inexperienced NES and advanced NNES raters. These results suggest that experienced NES raters were better at identifying NS norms than the other two rater groups.

Table 3. Mean, Standard Deviation, Effect Size (*d*), the Robust Welch-James ADF Pairwise Comparison (*F* statistic), and the Corresponding *p* Value by Rater Group

Rater group	Comparison	M_1	SD_1	M_2	SD_2	*d*	*F* statistic	*p* value
Inexperienced NES	**Native English speaker vs. child arrivals**	**8.56**	**0.28**	**6.01**	**1.62**	**1.99**	**52.33**	**< .001**
Inexperienced NES	**Child arrivals vs. adolescent arrivals**	**6.01**	**1.62**	**4.02**	**1.32**	**1.32**	**16.31**	**< .001**
Inexperienced NES	Adolescent arrivals vs. adult arrivals	4.02	1.32	3.24	1.06	0.65	2.96	.098
Experienced NES	**Native English speaker vs. child arrivals**	**8.96**	**0.06**	**7.35**	**1.26**	**1.64**	**36.37**	**< .001**
Experienced NES	**Child arrivals vs. adolescent arrivals**	**7.35**	**1.26**	**5.70**	**1.03**	**1.40**	**18.41**	**< .001**
Experienced NES	Adolescent arrivals vs. adult arrivals	5.70	1.03	4.85	1.24	0.75	3.92	.059
Advanced NNES	**Native English speaker vs. child arrivals**	**8.51**	**0.44**	**6.89**	**1.23**	**1.63**	**32.24**	**< .001**
Advanced NNES	**Child arrivals vs. adolescent arrivals**	**6.89**	**1.23**	**5.57**	**1.10**	**1.11**	**11.16**	**.002**
Advanced NNES	Adolescent arrivals vs. adult arrivals	5.57	1.10	5.14	1.11	0.39	1.08	.307

Note. M_1 and SD_1 as compared to M_2 and SD_2 are the means and standard deviations of ratings for the first and second speaker group in each comparison. Nonbold comparisons were not statistically significant after the Bonferroni adjustment ($\alpha = .05/21 = .002$) for Type I error.

Table 4. Mean and Standard Deviation (in parentheses) of the Three Rater Groups' Ratings for Each Speaker Group, along with the Robust Welch-James ADF Pairwise Comparison Results by Speaker Group

Speaker group	Advanced NNES	Inexperienced NES	Experienced NES	Comparison
Native English speaker	8.51 (0.44)	8.56 (0.28)	8.96 (0.06)	**Inexperienced NES vs. experienced NES (F = 37.28; p < .001; d = -2.03)**
				Experienced NES vs. advanced NNES (F = 16.06; p < .001; d = 1.19)
				Inexperienced NES vs. advanced NNES (F = 0.39; p = .542; d = 0.09)
Child arrivals	6.89 (1.23)	6.01 (1.62)	7.35 (1.26)	**Inexperienced NES vs. experienced NES (F = 125.95; p < .001; d = -0.72)**
				Experienced NES vs. advanced NNES (F = 12.82; p < .001; d = 0.37)
				Inexperienced NES vs. advanced NNES (F = 33.61; p < .001; d = -0.52)
Adolescent arrivals	5.57 (1.10)	4.02 (1.32)	5.70 (1.03)	**Inexperienced NES vs. experienced NES (F = 99.70; p < .001; d = -1.28)**
				Experienced NES vs. advanced NES (F = 0.96; p = .345; d = 0.12)
				Inexperienced NES vs. advanced NNES (F = 58.18; p < .001; d = -1.24)
Adult arrivals	5.14 (1.11)	3.24 (1.06)	4.85 (1.24)	**Inexperienced NES vs. experienced NES (F = 233.96; p < .001; d = -1.26)**
				Experienced NES vs. advanced NNES (F = 1.88; p = .193; d = -0.24)
				Inexperienced NES vs. advanced NNES (F = 115.02; p < .001; d = -1.73)

Note. Nonbold comparisons were not statistically significant after the Bonferroni adjustment (α = .05/21 = .002) for Type I error.

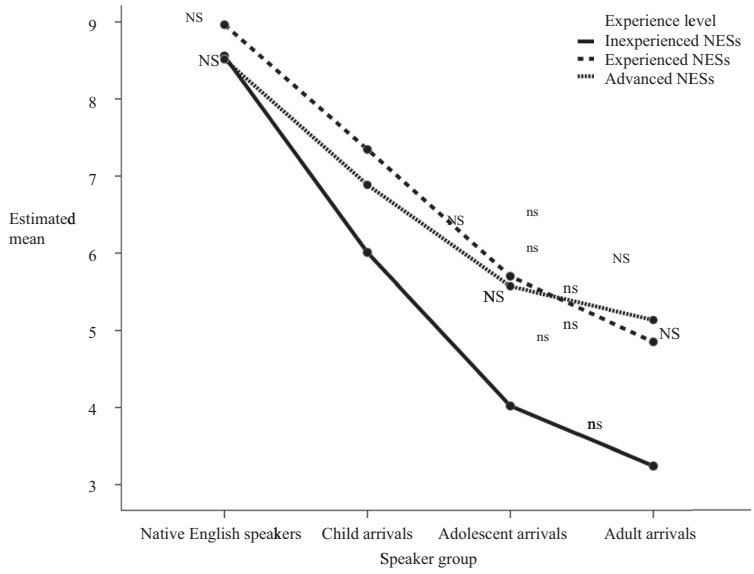

Figure 1. Estimated means from the ADF mixed ANOVA by rater group by speaker group. NS denotes nonsignificant within-group (rater group) mean differences, whereas ns represents nonsignificant between-groups (speaker group) mean differences, for those means nearest to the corresponding labels. All results are based on the Welch-James ADF pairwise comparison tests.

As seen in Figure 1 and Table 4, the following three nonsignificant rater group differences emerged: inexperienced NES versus advanced NNES raters at the NES speaker group and experienced NES versus advanced NNES raters at the adolescent and adult arrivals groups. For rating L2 speakers, whereas the experienced NES and advanced NNES raters had relatively consistent downward rating trends with the increase of L2 speakers' AoAs, these downward changes were much steeper for inexperienced NES raters. This implies that inexperienced NES raters tend to provide lower ratings, with these ratings decreasing at a greater rate than for the other two rater groups. Although slight differences existed between experienced NES and advanced NNES raters at times (see NES and child arrival comparisons), they were relatively small, especially for adolescent and adult arrivals.

To determine whether the three groups of raters identified the same L2 speakers as having achieved nativelike proficiency, we adopted the common nativelikeness criterion in the literature—that is, within two standard deviations of the average ratings assigned to NES controls—and we called the value derived from the criterion a *criterion score*. Judging from the ratings assigned to NES control speakers, experienced

NES raters appeared to be better able to distinguish between NSs and NNSs and assigned the highest ratings to NES control speakers of the three rater groups. Experienced NES raters also exhibited the least variation (M = 8.96, SD = 0.06); thus their criterion score was the highest among the three groups (8.84). Advanced NNES raters' mean ratings of NES control speakers (M = 8.51) were similar to those of inexperienced NES raters (M = 8.56) despite their nonnative background. However, there was less variation among inexperienced NES raters (SD = 0.28) than among advanced NNES raters (SD = 0.44). Advanced NNES raters' criterion score was thus the lowest (7.63).

We identified the most nativelike cases using advanced NNES raters' criterion score and the fewest cases by experienced NES raters' criterion score (see Table 5). Critically, no L2 speakers with an AoA of more than 5 were identified as nativelike by the experienced NES raters' criterion score, suggesting a potential terminus of plasticity at the age of 5. However, by the criterion score of the inexperienced NES rater group, two additional L2 speakers beyond the AoA of 5—one with an AoA of 6

Table 5. Nativelike Criteria and the Number of Nativelike L2 Speakers by Speaker Group by Rater Group

Nativelike criteria and number of nativelike L2 speakers	Advanced NNES Raters	Inexperienced NES Raters	Experienced NES Raters
Nativelike criterion score	7.63	8	8.84
Number of child arrivals identified as nativelike with AoAs ≤ 5 (initials and AoAs)	n = 1 YL, AoA = 5**	n = 1 YL, AoA = 5**	n = 1 YL, AoA = 5**
Number of child arrivals identified as nativelike with AoAs > 5 (initials and AoAs)	n = 7 LI, AoA = 6* SC, AoA = 8* MY, AoA = 7 AF, AoA = 7 KS, AoA = 8 AW, AoA = 9 SL, AoA = 10	n = 2 LI, AoA = 6* SC, AoA = 8*	n = 0
Number of adolescent arrivals speakers identified as nativelike	n = 0	n = 0	n = 0
Number of adult arrivals speakers identified as nativelike	n = 0	n = 0	n = 0

Note. ** = nativelike cases identified by all three rater groups; * = nativelike cases identified by two of the three rater group.

and the other of 8—met the nativelike criterion. According to the advanced NNES raters' criterion score, five L2 speakers with an AoA of more than 5 (including one child arrival with an AoA of 10) qualified as having obtained a nativelike accent.

DISCUSSION

The current study set out to address a methodological gap in the literature on age of learning effects on L2 production outcomes, specifically focusing on rater differences in the assessment of FAs. The study provided a systematic and detailed comparison of the judgments of three groups of raters with different L1 backgrounds and language experience. Using a 1–9 numerical scale, all raters judged the degrees of FAs in the paragraph read by 14 NES control participants and 50 L2 speakers with varying AoAs. To examine differences across rater groups in interrater reliability, two reliability measures were calculated for each rater group and were compared across groups. Approximate degrees of freedom mixed ANOVA analyses were conducted to examine the differences in raters' scaling of the speakers' FAs, the severity of ratings, and their ability to distinguish NESs and NNESs.

Results from the interrater reliability analysis showed that advanced NNES raters were the least reliable among the three rater groups, but the two NES rater groups exhibited similar degrees of rating consistency. These findings contradicted the results of Thompson's (1991) study, in which experienced NES raters were much more reliable than their inexperienced NES peers. Although a direct comparison of the current study with that of Thompson is not possible, we speculate that the discrepancies between the two may be attributed to the differences in the AoA ranges of the speakers, their level of proficiency in the L1, or the rating scales. The AoAs of the L2 speakers in Thompson's study ranged from 4 to 42, but the distribution of the AoA range was not specified. It could be that the majority of the speakers in Thompson's study were in a similar AoA range, which made it easy for raters to reach consensus on the speakers' FAs. In contrast, the current study included approximately equal numbers of participants in each AoA bracket. Furthermore, Thompson included only Russian learners of English who still maintained a high level of proficiency in their L1 and suggested that this may have resulted in heavier FAs. In contrast, the speakers in the present study self-reported substantial exposure to their L2 and may thus have lighter, and therefore harder-to-discern, FAs. Finally, the wider range of the scale (1–9) used in the current study may also lower the interrater reliability more so than the 1–5 scale in Thompson's study.

The results revealed no significant differences in raters' rank orders of the speakers' FAs; the three rater groups rank ordered the speakers

in a similar manner (NES > L2 child; L2 child > L2 adolescent; L2 adolescent = L2 adult). This finding replicated previous research (Derwing & Munro, 2013; Flege, 1988; Flege et al., 1997; MacKay et al., 2006) and suggests that raters can reliably evaluate the relative degrees of FA among speakers regardless of their own L1 backgrounds and language experience, even when the speech material is a paragraph rather than a single sentence.

Despite the similarities in their rank orders, the raters diverged in the severity of their ratings of L2 speakers' FAs. The analyses of their aggregated ratings revealed that, in general, inexperienced NES raters were consistently stricter (i.e., assigned lower ratings) than experienced NES raters and advanced NNES raters when judging L2 speakers' FAs. The discrepancies between experienced and inexperienced NES raters' severity corroborated previous research by Thompson (1991) but contradicted Bongaerts (1999b) and Bongaerts et al. (1997). The results were also consistent with previous research on oral language assessments showing that experienced NS raters were more lenient than inexperienced raters (Fayer & Krasinski, 1987; Hsieh, 2011) but contradicted the findings of Hadden (1991), which found the opposite effect.

The discrepancies between the three groups' severity in ratings may be attributed to a cognitive or exposure-based account, a social account, or a combination of both. Usage-based models emphasize the importance of experience and exposure in language acquisition (Kuhl, 2004; Tomasello, 1992); learners start with extracting phonemic prototypes or lexically specific schemas from the linguistic environment. Repeated exposure to the same prototypes or schemas then facilitates the development of phonetic categories and eventually mental grammar. Speech processing research (Bradlow & Bent, 2003, 2008) also provides some support for this account by demonstrating that, with exposure to foreign-accented speech in a controlled experimental session, listeners can adapt to nonnative speech and can improve intelligibility and comprehension of the accented speech. Due to prior language experience or exposure to FAs, experienced NES and advanced NNES raters in the current study may thus have developed phonetic categories for the accented speech, making it easier for them to process it. It is therefore possible that the higher ratings assigned by some raters reflected this ease of processing.

Furthermore, experienced NES and advanced NNES raters may also have been more lenient than inexperienced NES raters because they were familiar with a fuller range of FAs, having encountered L2 speakers with stronger FAs and more limited proficiency than those of the advanced L2 speakers in the current study (who were all long-term U.S. residents). Experienced NES and advanced NNES raters may have thus used the relatively lower proficiency L2 speakers from their prior experience as a baseline for comparison, a rating strategy documented in prior

studies (Isaacs & Thomson, 2013). It is worth noting that two recent studies have found that NSs' level of education and professional training influence their linguistic skills (specifically lexical knowledge and speaking proficiency; Mulder & Hulstijn, 2011) as well as their judgments about the structure of long-distance dependencies (Dąbrowska, 2010). These findings seem to corroborate the cognitive or exposure-based account beyond the domain of speech perception.

Alternatively, even if the required cognitive effort were equal, experienced NES and advanced NNES raters may have been more sympathetic with and favorable toward NNESs due to their familiarity with the NNESs' FAs (the advanced NNES raters were themselves NNESs). In other words, teaching experience and accent familiarity may have helped these raters recognize the NNESs' country of origin and ethnicity, or it may have even directly activated raters' favorable attitudes toward NNESs, leading to more lenient ratings. This can be illustrated in a qualitative study that investigated the effect of raters' familiarity with speakers' FAs on their rating process. Winke et al. (2013) asked heritage language speakers of Mandarin Chinese, Korean, and Spanish to rate ESL speech samples from speakers who had one of these three languages as their L1. The rating session was video recorded, and raters later watched these videos and discussed their decision-making processes. The heritage language speaker raters often claimed to recognize the L1s of some ESL speakers and reported that their ratings of them were more lenient as a result. To tease these two accounts apart, and to understand raters' cognitive processing and reactions to nonnative speech, future studies may utilize the sort of think-aloud protocols that are commonly employed to examine decision-making processes (e.g., Ericsson & Simon, 1980; Pressley & Afflerbach, 1995).

Although there were no significant differences between the aggregated ratings of experienced NES and advanced NNES raters in most pairwise comparisons in the current study, experienced NES raters were significantly more lenient when rating child L2 speakers compared to advanced NNES raters, but not reliably so when rating the adolescent and adult L2 speakers. The results suggest that raters' behavior varied by speakers' proficiency. We argue that our experienced NES raters' leniency may be attributed to their language experience as well. As previous research comparing teacher and nonteacher raters has shown, teachers are generally more lenient toward L2 learners than nonteachers (Fayer & Krasinski, 1987; Hsieh, 2011). Although the experienced NES raters in the current study did recognize the slight FAs in the child L2 speakers, their ratings may reflect knowledge of these speakers' likely improvement from some previous state.

Turning now to the results regarding the ability to distinguish between NESs and NNESs, experienced NES raters were, as a group, better at distinguishing native and nonnative speech, consistently assigning

the highest ratings to NESs. This particular finding is in line with what Flege et al.'s (1997) study revealed about the perceptual differences between NES raters in Alabama and those in Ottawa. It is likely that experienced NES raters' ESL or English as a foreign language teaching experience, combined with their familiarity with FAs and NES speech, facilitated the identification of NS norms. In line with the cognitive or exposure-based account, their experiences could have provided them with useful knowledge about variation in both NES and NNES speech, resulting in increased accuracy in distinguishing the two. In contrast, inexperienced NES raters have little experience with the speech variation in L2 speech, possibly leading them to interpret L1 variants as FAs (Markham, 1997). Therefore, advanced NNES raters and inexperienced NES raters did not differ reliably in their ratings of the NES speakers. We interpret these results as evidence supporting the hypothesis that NNSs, particularly those with substantial experience and high proficiency in the target language, can develop phonetic representations that are identical to *some* of those possessed by NSs of the target language. It is important to note that we are not arguing that NNSs can achieve nativelike mental representations of *all* of the segmental and suprasegmental features in the target language. In fact, the perceptions of FAs may involve potential paralinguistic cues such as speakers' confidence and both segmental and suprasegmental deviations (Piller, 2002). Research on crosslinguistic speech perception has shown that even early L2 learners with near-native proficiency in the target language may not achieve nativelike segmental perception (Navarra, Sebastián-Gallés, & Soto-Faraco, 2005; Pallier, Bosch, & Sebastián-Gallés, 1997; Sebastián-Gallés & Soto-Faraco, 1999). Nonetheless, corroborating findings on the perceptions of regional dialects and FAs (Baker, Eddington, & Nay, 2009; Flege et al., 1997), our results also demonstrate that NNSs can develop global sound representations of the target language resembling those of linguistically inexperienced NSs before developing the ability to produce these sounds in a targetlike way.

The discrepancies in the raters' ability to distinguish NS and NNS speakers resulted in different criteria for determining cases of nativelike L2 speakers. Because the experienced NES raters were better and more consistent as a group in identifying NSs, their nativelike criterion score was the highest among the three rater groups. Using their criterion, no L2 speakers with an AoA of more than 5 were identified as speaking with a nativelike accent. In contrast, several L2 speakers with an AoA of more than 5 met the nativelike criteria by both inexperienced NES raters and advanced NNES raters. One child L2 speaker with an AoA of 10 also qualified as a nativelike learner based on the advanced NNES raters' criterion score, though not by the other two NES rater groups' standards. The discrepancies in the number of nativelike cases can have important implications for the theories about age effects in L2

acquisition. Such discrepancies, particularly those between the two NES rater groups, would lead to different conclusions about the terminus of the critical period for L2 speech production. If we adopt experienced NES raters' criterion, the cutoff age would be around 5, which is close to Michael Long's (2005) proposal of age 6. However, using the other two rater groups' criteria would result in different conclusions because the upper limit of nativelike cases increased to 8 for inexperienced NES raters and 10 for advanced NNES raters.

The implications of the current study are thus multifold. First, the results suggest that, for future assessment tasks involving only rank ordering of speakers, advanced NNSs can be just as competent as NSs. Native speakers have long served as the norm for language instruction as well as assessment, and their evaluations have been regarded as more valid and informative than those of NNSs (Munro et al., 2006; Zhang & Elder, 2011). However, the current results show that the evaluations of NNSs with advanced proficiency are not reliably different from those of NSs in the scaling of L2 speakers' speech production. The NNS raters' degrees of severity in rating also closely paralleled those of experienced NS raters, except for the child L2 speaker group, which was judged less favorably by NNS raters than by the NS raters. In other words, the evaluations of advanced NNS raters resembled those of experienced NS raters in both rank ordering and overall severity. The results spell good news for the large number of NNS teachers worldwide and reassure that their evaluations and feedback to their L2 students are valid and meaningful.

Additionally, the discrepancies in the raters' severity, particularly those between experienced and inexperienced NS raters, have implications for the validity of research comparing L2 speakers' outcomes across studies. Recently, meta-analysis research has been gaining popularity in L2 acquisition research (Norris & Ortega, 2000), as it illuminates and advances the field by systematically synthesizing the evidence across empirical investigations. However, given the rater differences evident in the current study, future metasynthesis of age effect research would need to take into consideration the fact that such differences may constitute a confound. Inferences drawn from cross-study comparisons that fail to take rater backgrounds into account are therefore tenuous.

Finally, the existence of divergent rating outcomes between experienced and inexperienced NS raters also raises concerns about the *native speaker construct*. Native speakers have traditionally been defined as speakers of the standard variety of the language (Davies, 1999), but this definition lacks specificity. Researchers have challenged the absolute distinction between NSs and NNSs, arguing against such a simple dichotomy (Davies, 2003; Faez, 2011). As revealed in the current study, NSs' language and teaching experience, or lack thereof, may also influence their speech perception and judgments of NNSs' L2 proficiency.

The current conceptualization of NSs thus fails to appreciate the variation in L1 proficiency among even NSs themselves (Cook, 1999; Davies, 2003).

In addition to serving as the benchmark for comparison in L2 acquisition research, NS norms have also been used widely in language teaching and testing. The lack of detail in the current definition, however, raises questions about who the ideal NSs actually are. Would simultaneous bilinguals (those who have been exposed to two languages since birth) of the standard variety of both of their languages be regarded as NSs of both languages? Would second-generation immigrants whose exposure to the target language began at school age but who have become dominant in the target language be considered NSs? If not, would the construct of NSs and NNSs become one of monolinguals and bilinguals? Although we agree on the necessity of concrete and well-defined benchmarks in empirical research, we urge researchers to acknowledge the complexities of the NS construct, to provide justifications for their decisions, and to include detailed descriptions of any raters (NSs or NNSs) their study employs. In a review of factors that affect L2 FAs, Piske, MacKay, and Flege (2001) discussed the potential rater effect in the literature and proposed that researchers include a representative sample of raters from different backgrounds rather than one particular type. The proposal appeared to be a feasible and balanced solution. However, we would like to suggest that researchers also evaluate the purpose of their own study before making a decision on the specific type of raters to include. For example, if the purpose of the study is to evaluate the L2 speakers' FAs as perceived by NSs living in metropolitan cities (who are presumably familiar with FAs), then it would make sense to include judgments by experienced NSs rather than those of inexperienced NSs. Conversely, if the intent is to assess L2 speakers' FAs as perceived by NSs with limited exposure to L2 speakers, using inexperienced NSs would provide more informative assessments than would using experienced NSs.

We suggest several directions for future research. First, although the size of the current rater sample (n = 30) is reasonable and relatively large for studies on age effects and experimental or quasi-experimental L2 acquisition research in general, it is not sufficient for more sophisticated statistical techniques that also take into account variation on the speaker level, such as multifaceted Rasch models (Hsieh, 2011; Isaacs & Thomson, 2013; Winke et al., 2013) or G-theory (Xi & Mollaun, 2011). Future researchers may want to include a larger sample and to utilize such advanced analyses. Additionally, in the present study, we included only one type of speech stimuli—namely, the production of a read-aloud passage. Previous research has also shown that speakers' proficiency may vary depending on the type of L2 tasks under study (Moyer, 2004; Thompson, 1991); future work investigating the effect of task types will advance our understanding of the matter. Finally, further

research is also needed to better understand the roles of other rater background variables, such as gender and attitudes toward NNSs, and to assess the interaction between such variables and task types.

CONCLUSIONS

To conclude, the current study found that L1 background and language experience significantly influenced raters' rating scores and ability to distinguish NSs from near-native early L2 learners. However, such experience was not found to predict raters' scaling of L2 speakers' proficiency. Results from the current study therefore demonstrate that differences among raters have important implications for studying age of learning effects and should be controlled for in this line of research. The results also indicate that rater differences can affect the number of nativelike L2 learners identified. Although no prior research has yet adapted NNSs' ratings as criterion scores to determine the cutoff age for the critical period, NNSs have been included as raters in several studies investigating age effects on L2 acquisition (Flege, 1988; MacKay et al., 2006). Additionally, as discussed in the literature review section, information about the experience level of NS raters included in age effect research is generally very limited. The current study revealed variation among NS raters' judgments, which can potentially lead to different conclusions about the terminus of the critical period. This particular finding thus highlights critical implications for theories of L2 acquisition.

Rater differences also impact the severity of ratings, rendering comparisons or synthesis across studies problematic. In addition, the findings raise questions about the general reliability and validity of instruments used in measuring L2 outcomes. We hope that this study will draw researchers' attention to the methodological issues in L2 acquisition research and that more research will be conducted to help ensure the reliability and validity of the inferences informing L2 acquisition theories.

Received 23 July 2013
Accepted 10 April 2014
Final Version Received 20 June 2014

NOTES

1. Previous research used either *listeners* or *raters* depending on the purpose of their studies. We chose to use *raters* instead of *listeners* to denote active evaluation, rather than passive listening, on the part of the participants.

2. We are aware of debates on the critical period hypothesis, which is not accepted by all researchers as a useful account for the age of learning effect. For example, rather than attributing the age effect to a critical period, Flege (1995) proposed the speech learning model to explain age effects on the acquisition of L2 segments. He argued that L1 AQ3

and L2 sounds share the same phonological space and mutually influence each other. Because adults have more robust L1 categories compared to children, it is more challenging for adults to form new L2 categories. However, because the focus of this study is on rater effects (a measurement issue) in age of learning research rather than on the critical period hypothesis per se, we only discuss empirical studies that are relevant to the current study.

3. One reviewer brought to our attention an additional study by Bongaerts (Bongaerts, 1999a) that involved nonnative speaker judges. This particular study, which was written in Dutch, is not included in the current article because we limit our literature review to publications in English only.

4. We use *scale* and *rank* interchangeably throughout the article.

5. Note that here we limit our discussion to empirical studies that examined the age of learning effect. We understand that the distinction we made between rater studies in age effect research and in other research may be arbitrary. However, the distinction is essential to the aim of the current article, which is to raise awareness of an important methodological issue in age of learning effect research and in L2 acquisition research generally.

6. We decided to include speech therapists because speech therapists, along with phoneticians and ESL teachers, have been referred to as experienced or "expert" raters in L2 pronunciation research (Isaacs & Thomson, 2013). Speech therapists may develop familiarity with foreign accents through their work in accent modification or reduction (see http://www.asha.org/public/speech/development/accent-modification/). There was only one speech therapist in the experienced NES rater group. The other nine raters in the group were current or past ESL teachers. The speech therapist met the criteria for inclusion, as did the other nine raters in the group. **AQ4**

7. We conducted a t test comparing the aggregated ratings of the three experienced NES raters who speak Mandarin and the seven who do not, and the differences were not significant, $t(8) = -.433$, $p = .66$. On the basis of these results, we do not differentiate the two subgroups in our analyses.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*, 249–306.

Asher, J. J., & García, R. (1969). The optimal age to learn a foreign language. *Modern Language Journal*, *53*, 334–341.

Audacity Team. (2000). Audacity (Version 1.2.5) [Computer software]. Retrieved from http://audacity.sourceforge.net/

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*, 238–257.

Baker, W., Eddington, D., & Nay, L. (2009). Dialect identification: The effects of region of origin and amount of experience. *American Speech*, *84*, 48–71.

Barnwell, D. (1989). "Naïve" native speakers and judgments of oral proficiency in Spanish. *Language Testing*, *6*, 152–163.

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 5.1.05) [Computer program]. Retrieved from http://www.fon.hum.uva.nl/praat/

Bongaerts, T. (1999a). De keuze van beoordelaars in onderzoek naar uitspreekvaardigheid in een vreemde taal [Selecting raters for research on L2 pronunciation]. In M. Gerritsen & D. Springorum (Eds.), *Een bundel bedrijfscommunicatie voor Ger Peerbooms bij gelegenheid van zijn 65e verjaardag* (pp. 1–10). Nijmegen, the Netherlands: Nijmegen University Press.

Bongaerts, T. (1999b). Ultimate attainment in foreign language pronunciation: The case of very advanced late foreign language learners. In D. Birdsong (Ed.), *Second language acquisition and the Critical Period Hypothesis* (pp. 133–159). Mahwah, NJ: Erlbaum.

Bongaerts, T., Mennen, S., & van der Slik, F. (2000). Authenticity of pronunciation in naturalistic second language acquisition: The case of very advanced late learners of Dutch as a second language. *Studia linguistica*, *54*, 298–308.

Bongaerts, T., Planken, B., & Schils, E. (1995). Can late starters attain a native accent in a foreign language? A test of the Critical Period Hypothesis. In D. Singleton & Z. Lengyel (Eds.), *The age factor in second language acquisition* (pp. 30–50). Clevedon, UK: Multilingual Matters.

Bongaerts, T., van Summeren, C., Planken, B., & Schils, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *Studies in Second Language Acquisition*, *19*, 447–465.

Bradlow, A., & Bent, T. (2003). Listener adaptation to foreign-accented English. In M. J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1581–1583). Barcelona, Spain: Universitat Autònoma de Barcelona.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*, 707–729.

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility. *Language Learning*, *22*, 203–220.

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second Language Studies*, *21*, 1–43.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*, 201–219.

Chalhoub-Deville, M. (1995). A contextualized approach to describing oral language proficiency. *Language Learning*, *45*, 251–281.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, *116*, 3647–3658.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, *33*, 185–209.

Dąbrowska, E. (2010). Naive vs. expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, *27*, 1–23.

Davies, A. (1999). Standard English: Discordant voices. *World Englishes*, *18*, 171–186.

Davies, A. (2003). *The native speaker: Myth and reality*. Clevedon, UK: Multilingual Matters.

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*, 163–185.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.

Faez, F. (2011). Reconceptualizing the native/nonnative speaker dichotomy. *Journal of Language, Identity & Education*, *10*, 231–249.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, *37*, 313–326.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America*, *84*, 70–79.

Flege, J. E., Frieda, E. M., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics*, *25*, 169–186.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, *97*, 3125–3135.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, *41*, 78–104.

Galloway, V. B. (1980). Perceptions of the communicative efforts of American students of Spanish. *Modern Language Journal*, *64*, 428–433.

Gass, S., & Varonis, E. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*, 65–89.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second-language communication. *Language Learning*, *41*, 1–24.

Hsieh, C.-N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *9*, 47–74.

Huang, B. H. (2014). The effects of age on second language grammar and speech production. *Journal of Psycholinguistic Research*, *43*, 397–420.

Huang, B. H., & Jun, S.-A. (2011). Specifying the age-related effect on the acquisition of second language prosody. *Language and Speech*, *54*, 387–414.
Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*, 135–159.
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*, 459–489.
Keselman, H. J., Algina, J., & Kowalchuk, R. K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, *54*, 1–20.
Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187–217.
Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, *5*, 831–843.
Kunnan, A. J. (Ed.). (2000). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (Vol. 9). Cambridge, UK: Cambridge University Press.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin*, *117*, 547–560.
Long, M. H. (2005). Problems with supposed counter-evidence to the Critical Period Hypothesis. *International Review of Applied Linguistics*, *43*, 287–317.
MacKay, I. R., Flege, J. E., & Imai, S. (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics*, *27*, 157–183.
Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, *29*, 539–556.
Markham, D. (1997). *Phonetic imitation, accent, and the learner*. Lund, Sweden: Lund University Press.
MATLAB (Version 6.5.1) [Computer software]. Natick, MA: MathWorks.
Moyer, A. (1999). Ultimate attainment in L2 phonology. *Studies in Second Language Acquisition*, *21*, 81–108.
Moyer, A. (2004). *Age, accent and experience in second language acquisition*. Clevedon, UK: Multilingual Matters.
Mulder, K., & Hulstijn, J. H. (2011). Linguistic skills of adult native speakers, as a function of age and level of education. *Applied Linguistics*, *32*, 475–494.
Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition*, *28*, 111–131.
Munro, M., & Mann, V. (2005). Age of immersion as a predictor of foreign accent. *Applied Psycholinguistics*, *26*, 311–341.
Navarra, J., Sebastián-Gallés, N., & Soto-Faraco, S. (2005). The perception of second language sounds in early bilinguals: New evidence from an implicit measure. *Journal of Experimental Psychology: Human Perception and Performance*, *31*, 912–918.
Nikolov, M. (2000). The Critical Period Hypothesis reconsidered: Successful adult learners of Hungarian and English. *International Review of Applied Linguistics in Language (IRAL) Teaching*, *38*, 109–124.
Norris, J. M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.
Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*, 261–283.
Pallier, C., Bosch, L., & Sebastián-Gallés, N. (1997). A limit on behavioral plasticity in speech perception. *Cognition*, *64*, B9–B17.
Piller, I. (2002). Passing for a native speaker: Identity and success in second language learning. *Journal of Sociolinguistics*, *6*, 179–208.
Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*, 191–215.
Pressley, M., & Afflerbach, P. P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.
Scovel, T. (2000). A critical review of the critical period research. *Annual Review of Applied Linguistics*, *20*, 213–223.

Sebastián-Gallés, N., & Soto-Faraco, S. (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition*, *72*, 111–123.
Thompson, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, *41*, 177–204.
Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge, UK: Cambridge University Press.
Trofimovich, P., Lightbown, P. M., Halter, R., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition*, *31*, 609–639.
Urbaniak, G. C., & Plous, S. (2013). Research randomizer (Version 4.0) [Computer software]. Retrieved from http://www.randomizer.org/
Weinberger, S. (2014). *Speech accent archive* [Database]. Retrieved from http://accent.gmu.edu.
Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly*, *47*, 762–789.
Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231–252.
Xi, X., & Mollaun, P. (2011). Using raters from India to score a large test. *Language Learning*, *61*, 1222–1255.
Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*, 31–50.

APPENDIX

READING PARAGRAPH FOR SPEECH ELICITATION

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

Author Queries

QA	The distinction between surnames can be ambiguous, therefore to ensure accurate tagging for indexing purposes online (eg for PubMed entries), please check that the highlighted surnames have been correctly identified, that all names are in the correct order and spelt correctly.
AQ1	Note that note callouts should not be placed in headings. I have moved the note 2 callout to the sentence below.
AQ2	Dunlap et al. 1996 is not listed in the Refs. Add to Refs.
AQ3	Flege 1995 is not listed in the Refs. Change to “Flege et al. (1995)” or add Flege 1995 to Refs.
AQ4	I have switched notes 5 and 6 b/c the note 5 callout seemed to correspond to the text of (former) note 6, and the note 6 callout matched (former) note 5.