What Does More Time Buy You? Another Look at the Effects of Long-Term Residence on Production Accuracy of English /ɹ/ and /l/ by Japanese Speakers

Jenifer Larson-Hall

University of North Texas, Denton, Texas

Key words

input

Japanese

length of residence

phonology

Abstract

This study tested the issue of whether extended length of residence (LOR) in adulthood can provide sufficient input to overcome age effects. The study replicates Flege, Takagi, and Mann (1995), which found that 10 out of 12 Japanese learners of English with extensive residence (12 years or more) produced liquids as accurately as native speakers of English (NS). Further, for both accuracy and native-like accentedness, the Japanese with extensive residence performed statistically better as a group than inexperienced Japanese (less than 3 years of residence). Results with a new sample of Japanese learners in this study found no statistical difference between the Japanese groups with extended versus short LOR although both reported equal levels of daily input in English. Additionally, both groups received statistically lower scores than NS. Moreover, LOR affected the two groups differently: The accuracy and native-like accentedness of words and sentences by Japanese with extensive residence declined with LOR (and chronological age when age of arrival was partialled out), while for Japanese with short residence accent improved with increased LOR (but not age). This study is the first to document a decline in second language production ability with LOR and age in older second language learners. However, this finding deserves to be addressed with further research, as the study was not designed to investigate this question and thus not all relevant factors, such as motivation or attitude, were controlled for. The results from the short-residence learners indicate that the initial one to two years of immersion may be the most important for improving phonological ability.

Acknowledgments: Thanks go to Alan Juffs, Robert DeKeyser, and Heather Goad, and several anonymous reviewers whose insightful comments greatly improved this paper. I also thank the University of Pittsburgh for financial assistance in conducting the study. My sincerest gratitude to Rich Herrington who helped with statistics and introduced me to the R statistical program. Keiko Rushlander was invaluable for helping find Japanese participants for this study. A preliminary version of this study was presented at the Japan Second Language Association Conference, Kyoto Sango University, Kyoto, Japan, May 18, 2002.

Address for correspondence. University of North Texas, Dept. of English, P.O. Box 311307, Denton, TX 76205; e-mail: <jenifer@unt.edu>.

1 Introduction

Recently there has been renewed interest in the role that large amounts of input can play in helping learners to improve their phonological accuracy. Although all large-scale studies of long-term immigrants have found that age of arrival is by far the most significant factor in explaining phonological performance (Flege, Munro, & MacKay, 1995; Flege, Yeni-Komshian, & Liu, 1999; Oyama, 1976), some researchers claim that age is highly interrelated to factors such as type and amount of input, and that these and other nonbiological factors are the underlying reason for age effects (Flege, 1987; Jia & Aaronson, 2003; Jia, Aaronson, & Wu, 2002; Moyer, 2004). If it is the quantity and / or quality of input itself that is important, we would expect that learners would continue to progress regardless of age as they receive further input. This study, therefore, seeks to examine whether long-term residence in a target-language country, with the accompanying implication of large amounts of input, is sufficient to show a relationship with continued improvement in phonological accuracy.

Flege and his colleagues have found that quantity of input, measured by length of residence (LOR) in a target-language country, makes a difference when comparing learners with very short and long LOR. Flege and Fletcher (1992) found a significant correlation with degree of accent on English sentences by Spanish L1 adults with an LOR of 14.3 years versus 0.7 years in the U.S. Flege, Bohn, and Jang (1997) found that learner groups with German, Spanish, Mandarin, and Korean L1 and an average LOR of 7.3 years performed better on a vowel production and perception task than learners with an average LOR of 0.7 years. However, in these cases the learners with a long LOR did not perform at the native speaker level.

For this reason, the study by Flege, Takagi, and Mann (1995) (hereafter FTM) stands out. FTM found that Japanese learners of English with an average LOR of 20.8 years not only performed better than the learners with an LOR of 1.6 years, but were indistinguishable from native speakers (NS) on a production test of words beginning with /ɹ/ and /l/. The findings of FTM (1995) bear on the Critical Period (CP) debate because they seem to show that increased input can result in native-like ability to produce sounds regardless of age of arrival (AOA). Evidence that adult learners could, with sufficient input, progress to native speaker levels, even if this took longer than for younger learners, would certainly indicate that the hypothesis of a CP was wrong.

Although more recent publications by Flege and his colleagues have focused on the continued use of the first language (L1) as a predictive factor in ultimate second language (L2) phonological attainment (Flege & MacKay, 2004; Flege, MacKay, & Piske, 2002; Guion, Flege, & Loftin, 2000; Piske, Flege, MacKay, & Meador, 2002; Piske, MacKay, & Flege, 2001), these researchers continue to build upon the research of the past and have not discarded the hypothesis that the quality and quantity of input may be the underlying cause of age effects. For example, even though Piske et al. (2001) conclude from a review of the literature that LOR exerts its strongest effect in the early stages of learning the L2, they nevertheless claim that "an effect of LOR is much more likely to be found if the groups of L2 learners examined differ greatly in terms of mean LOR values" (p. 199). Flege and Liu (2001) examines Chinese learners of English with a mean LOR of five years and concludes that “adults’ performance in an L2 will improve measurably over time, but only if they receive a substantial amount of native speaker input” (p. 527). Flege and MacKay (2004) explain that Flege’s speech learning model hypothesizes that, over time, adults have the capacity to “perceive the properties of L2 speech sounds accurately and to establish new phonetic categories” (p. 5).

1.1
The importance of length of residence

The present study questions whether extensive length of residence (LOR) can generally be expected to improve learners’ phonology in the way that Flege’s SLM hypothesizes and which FTM found in their study. When researchers measure LOR, they seek to do so in order to provide a quantifiable measure of the amount of input that learners have received. Of course researchers realize that simple residence in a country will not always provide intensive input to the learner. Jia and Aaronson (2003) demonstrated that younger Chinese immigrants to the U.S. (age 9 and below) tended to have more English-speaking friends, read for pleasure more in English, and preferred speaking English more than older children (age 12 and above). These differences surely lead to increased amounts of input for the younger learners. Likewise, several studies have found that as the second language (L2) proficiency of immigrants increases, the first language (L1) proficiency tends to decrease (Flege, Munro & MacKay, 1995; Jia et al., 2002; Yeni-Komshian, Flege, & Liu, 2000). The fact that younger learners tend to become L2 dominant while older learners remain L1 dominant would seem to imply that younger learners (who are more proficient) receive more L2 than L1 input.

However, if amount of input, motivation, and willingness to speak the language are some of the main factors behind the surrogate factor of biological age, why should we find such clear age-related patterns among child learners? That is, wouldn’t young immigrants of various ages evince a normal distribution of such choices as who to have as friends, whether to read in the L1 or L2, and whether to favor the L1 or L2 if these underlay the age effect rather than resulting from them? Instead, these factors seem to be results that flow from following the path of least resistance, given age effects. For example, Jia and Aaronson (2003) found that while younger children associated with children who lived nearby (who spoke English), the older children sought out Chinese-speaking friends who lived further away. However, in the case of TV, where it was perhaps nearly impossible to find Chinese-language programs, by the third year all of the older children watched English TV almost 100% of the time, even though they did not speak English at home and rarely read English books for pleasure. The older learners only followed the path of greater resistance (for them, watching English instead of Chinese TV) when they had little choice.

Among adult learners of a second language, individual differences in levels of motivation, attitudes towards the language and its speakers, and inherent language learning ability will play a much bigger role. Moyer (2004) mentions other factors that affect differential success among learners who have all started learning a language at an older age, such as amount and type of instruction, pronunciation training and identity affiliation. Dörnyei (2005) also mentions personality as a factor that most certainly affects language learning. A large-scale questionnaire study recently conducted by McManus and Pleines (2005) found that personality factors affect motivations and barriers to language learning.

Hyltenstam and Abrahamsson's (2003) consensus model of ultimate attainment allots a large role to social / psychological effects among learners after childhood but only a small role to such effects before adolescence.

Certainly there are many factors involved in second language success. However, LOR has not generally been taken to be one of the important ones for adults, beyond gains that may occur after initial immersion in the target language. Critical Period studies examining ultimate outcomes of child versus adult starters have not found LOR to play a significant role in explaining variance for adults who have resided in the target-language country for five years or more (Flege, 1999; Flege, Munro, & MacKay, 1995; see also DeKeyser & Larson-Hall, 2005, for a recent review of CP studies). In fact, a recent study by Moyer (2004), which asked many detailed questions about amount of input (years of formal instruction, the consistency of instruction, hours spent with NS per week, the consistency of such contact, and hours spent per week in passive activities like watching TV, reading, or writing) found that none of these experimental variables reached significant levels of correlation with performance on a variety of phonological tasks (when 2 early bilingual outliers were factored out).[1] The results of FTM are thus quite surprising, given previous research, and deserve a second look.

Although clearly LOR is only an indirect measure of actual amount of input, the purpose of the present research is not to deal directly with the issue of the validity of LOR as a substitute for input, but to compare relative performance by means of this previously used criterion — however flawed it may be.

1.2 *Monitoring pronunciation*

An impressive finding from FTM (1995) was that the group with a long LOR did as well as NS on three different kinds of tasks: elicited production of isolated words, reading a word-list, and spontaneous production of a sentence of their own choosing using the word in question. This seems to indicate that the experienced Japanese learners were able to function well in more spontaneous as well as more constrained tasks. However, in all three tasks the participants were presumably aware of the focus on the /ɹ/ and /l/ distinction, and thus they may have been monitoring all of their productions. In the present study the question of whether learners' monitoring of production would play any role in accuracy was addressed by asking participants to read aloud a story in which the /ɹ/ and /l/ words were embedded before they read a word list. Although the story reading task could not guarantee that learners would not focus specifically on the contrast, it was presumed that reading at a natural rate would naturally lead the learners to focus more on meaning, and also leave them less opportunity to monitor their pronunciation.

[1] Moyer did find a significant ($p = .0009$, $r = .67$) correlation between LOR and accent scores, but as this was not a comparison between child and adult learners who had reached final developmental levels, this finding should not be compared to previously mentioned CPH studies.

Second language acquisition (SLA) studies which have focused on attention to task or monitoring have found inconsistent results. We might expect that learners would produce sounds more accurately when focusing on form (reading a word list, sentences or a text) rather than meaning (speaking spontaneously). Skehan and Foster (2001), although admittedly focusing on morphosyntax, not phonology, have argued that fluency and form have been shown to load on separate factors in a factor analysis of speech, meaning that if fluency is high, accuracy of form is generally low, and vice versa. In the phonological realm, some studies have found better accuracy in spontaneous speaking tasks than in sentence reading (Moyer, 2004; Oyama, 1976; Sato, 1985; Thompson, 1991). This seems to be the case when the sentences contain many difficult sounds, whereas in spontaneous speech the participants can to some extent avoid a concentrated string of difficult sounds. On the other hand, other studies have found greater accuracy in list or sentence reading than in spontaneous production (Bongaerts, Planken, & Schils, 1995; Dickerson & Dickerson, 1977; Major, 1994). The problem with comparing results on reading tasks versus spontaneous speech is that the materials are always different, and it is unknown to what extent this may also affect accent scores. Bongaerts et al. (1995) found that nonexceptional Dutch learners of English performed better on a text-length reading than on sentence reading, although they still performed best on word-list reading.

Given the available research, the hypothesis here is that all Japanese learners, regardless of their LOR, will perform better on the word list task than the story-reading task, while NS will perform equally on both tasks. Another task, not found in FTM, was included in this study. In this task the participants' accent on an entire sentence was rated. This task was used see how participants would be rated on a longer passage, since very good learners have been found to pass for NS on sentence-length productions (Bongaerts, 1999; Bongaerts, Mennen, & van der Slik, 2000).

1.3 *Research questions*

The present study addresses the following questions:

1. Will LOR correlate positively with production accuracy on phonological measures?
2. Is a long LOR (with the assumption of large quantities of input) sufficient to allow adult Japanese learners to improve their phonology to the level of NS on production accuracy and accentedness of words beginning with English /ɹ/ and /l/?
3. Will more monitoring of form result in more accurate production for NNS (pertinent to Experiment 1 only)?

2 Methods

2.1 *Participants*

Japanese speakers who had not previously lived in an English-speaking country before their stay in the U.S. participated in this study. There were 15 speakers in each of two groups: Inexperienced Japanese (IJ), who had lived in the U.S. three years or less, and Experienced Japanese (EJ), who had lived in the U.S. twelve years or more, replicating the conditions of FTM (1995) with three more participants in each group. Participants were recruited solely based on their LOR, without regard to proficiency, use of English, age of arrival, or any other factor besides LOR and availability. Participants were recruited through flyers hung at a university and networking with a Japanese association in the city. Note that (as is the case in most social science research, according to Klein, 2004, and as was most probably the case in FTM), participants were not randomly recruited. Furthermore, this study was nonexperimental in design (as in FTM, no treatment was administered). Thus, it is unclear to what extent results of this study (or of FTM) are generalizable to other populations of Japanese adults living in the U.S. One motive for replication is to determine whether FTM's results are indeed generalizable to another population.

Native English (NS) speakers (*n* = 15) served as a comparison group. Additionally, 12 native speakers, college undergraduates, served as judges of the production data (3 M, 9 F, *M* = 23.3yrs). Summary data for the speaking participants is found in Table 1. Individual measures for age of arrival, age at time of testing, length of residence, use of English, pronunciation instruction and scores on experiments are given in Appendix 1. Per FTM and the large-scale CPH studies, age of arrival was operationalized as the age at which participants first arrived in an English-speaking country for purposes of residing there, even though most participants probably had learned some English before arriving in the U.S.

Table 1

Characteristics of the two Japanese learners and one native speaker group

	Group					
	IJ (*N* = 15) M = 4, F = 11		**EJ (*N* = 15) M = 1, F = 14**		**NS (*N* = 15) M = 6, F = 9**	
	M *(SD)*	Range	M *(SD)*	Range	M *(SD)*	Range
Testing Age (years)	29.1 (4.2)	20–36	52.5 (8.4)	38–66	34.5 (8.0)	22–54
Arrival age (years)	27.1 (5.2)	17–33	28.7 (5.7)	22–38		
U.S. residence (years)	1.1 (0.7)	.3–2.8	23.2 (8.5)	12–42		
Ave. use of English[a] possible range 1–21	11.5 (4.9)	4–20	12.7 (5.7)	5–21		
Amt. of /ɹ/ and /l/ pron. instruction (hours)	7.8 (21.5)	0–84	3.5 (7.6)	0–24		

[a] The question for this measure was taken directly from FTM and was a response to the question: " On average, how often do you use English (circle one)," with each of the three areas — at home, at work, and in social settings — followed by a scale from 1 (never) to 7 (frequently).

As is to be expected, pairwise comparisons found a significant difference between the Japanese groups for testing age ($p < .0005$). Note that both groups of Japanese came to the U.S. at roughly the same mean age, however (there is no significant difference, $p = .91$). As per FTM (1995),the average use of English on the part of the speakers is reported. Note, however, that this is a measure of output, not input. Although input and output are most likely correlated, it would be quite possible for a participant to receive large amounts of English input but produce little English output. The IJ speakers reported using English slightly less on an average daily basis than the EJ, but the difference was not significant, ($p = .54$). As another crude measure of input, marriage to a NS was also ascertained. Eight of the EJ and one of the IJ were married to native speakers of English. The IJ had more hours of pronunciation training than the EJ but this was not a significant difference between the groups ($p = .72$).

2.2
Design and procedure

All speakers were recorded and tested individually in a quiet room, which was generally at their own home or office for the experienced Japanese, and a room at the University of Pittsburgh for the inexperienced Japanese. Each Japanese participant began the experiment by filling out a preliminary language background questionnaire. The participants took an aptitude test (which is not reported on here), then read the story and subsequently a word list of minimal pairs. The reading tasks were the only parts that the NS controls completed. Both reading tasks were recorded on a JVC or Tascam DAT recorder. After the recording section, participants completed an additional questionnaire about their pronunciation training with /ɹ/ and /l/. The judges listened to the productions in a quiet room on a computer using Sony MDR-CD780 headphones.

The test took one hour for the Japanese speakers, and payment of $10 for the IJ and $15 for the EJ was offered. The reading task took about 10 mins for NS speakers, who were offered no payment. The judging took one hour and all judges were paid $15. All judges reported normal hearing.

2.2.1
Reading tasks

All reading tasks were recorded in the author's presence. Participants were told that their pronunciation was being tested, but they were not specifically told about the focus on /ɹ/ and /l/. Participants read the story aloud once (see Appendix 2). They were not allowed to read through the story before recording it, as it was felt that they might pick up on the focus on /ɹ/ and /l/ if they had a chance to read the story first, and the main reason for the story was to obtain more natural tokens of the words beginning in /ɹ/ and /l/. The word list made it clear, however, that the focus was on /ɹ/ and /l/ (see Appendix 3). No distractor words were used as FTM did not use any, and use of any distractors might have been perceived as a reason for a difference between the two studies. Participants said each word in the word list twice, once in a carrier phrase (*The next word is* ____) and then again in isolation. In the case of the story participants were told to read at a normal pace, and that if they mispronounced a word they could correct themselves. For the word list, the participants were also told they could repeat the word if they made a mistake. If the participants mispronounced a word so that it did not rhyme with its minimal pair (such as [ɹuk] vs. [lʌk]), they were corrected by the experimenter and asked to repeat the word.

3 Experiment 1

The purpose of this task was to see how accurately words beginning in /ɹ/ and /l/ spoken by groups of Japanese learners differing in their LOR would be identified by native speakers of English. Two tasks were included to ascertain whether the amount of attention given to form would affect production accuracy.

3.1 *Procedure*

Eight minimal pairs (out of 20 possible) were taken from the story-reading and word-list task for judging. The words chosen are reproduced in bold in Appendix 3. To further reduce the judge's task, each judge listened to only 15 out of the 45 speakers, five from each of the three categories—IJ, EJ, and NS. Thus, in this experiment each judge listened to (15 speakers × 16 words × 2 conditions) = 480 utterances. All participants were judged by at least three judges (this is within the normal range of judges, according to Piske et al., 2001).

The listeners identified the initial consonant as /ɹ/ and /l/ and also indicated their degree of confidence in their forced-choice judgments. All stimuli were randomized through use of a computer program.[2] The judges heard a word through headphones and on the computer screen saw both members of the minimal pair (such as 'rip' and 'lip'). The judges used a mouse to click on one of four alternatives: (1) Definitely L, (2) Possibly L, (3) Possibly R, and (4) Definitely R. The judges were told that if they were fairly sure they should choose 'Definitely'. This 4-choice scale was used instead of the six-choice scale (Definitely, Probably, Possibly) used by FTM (1995) because in Experiment 1 of that study the judges were not consistent in their judgments—as a group, they did not judge the NS speakers to be better even than the IJ speakers.[3] With the four-point scale it was thought that some degree of confidence in the choice could be measured with less confusion than was afforded by

[2] Developed by Ammon B. Larson using Microsoft Visual Basic software and available upon request.

[3] Usually, participants' scores on a phonological measure are calculated by averaging the results of several judges and several words or sentences into one mean score. In the case of FTM's listener-based score, the focus was not on the score of each participant, but on the score each judge gave on the average to the NS, IJ, and EJ groups on the two different sounds. There were 10 judges in FTM, so that each judge had six scores (2 liquid consonants × 3 groups). These 60 listener-based scores (10 judges × 6 scores each) were subjected to an ANOVA to see if there were differences in the way the judges rated each of the three groups. If judges had rated the participants differently, varying as to their group affiliation, the ANOVA would have found a significant effect of the factor of group. However, it did not. This means, as FTM state, "the listeners were not sufficiently consistent among themselves to support the conclusion that NS speakers produced /ɹ/ and /l/ more accurately than even the IJ speakers" (1995, p. 33). FTM subsequently performed a weighted analysis of scores in order to compensate for this problem.

the six-choice scale. Although it is true that some researchers have argued for expanded scales (Chaudron, 1983; Flege & Fletcher, 1992), in the case of identification and not accent it seems clear from FTM (1995) that too many choices can be overwhelming for judges. The judges completed a trial use of the R/L identification task with 12 words from three speakers before they began the actual test.

3.1.1 Stimulus preparation

All of the productions were downloaded onto a computer and digitized at 48.0kHz with 16-bit resolution, then normalized for peak intensity. Words extracted from the story task were cut in the following manner: The cursor was placed at a point after which the /ɹ/ and /l/ sound could be heard, then moved left by 5ms steps. When any other sound before the initial /ɹ/ and /l/ of the intended word could be heard, the author went to the previous step and extracted the word from that point. Generally, the second repetition of the word from the reading list was chosen. However, as the author is a native speaker of English, if it was felt that the first token sounded more native-like, that one was chosen instead. No correction was made to control for different intonational contours between words extracted from the reading task and those taken from the word list task as it was not anticipated that this would affect word identification.

3.2 *Results*

3.2.1 Group comparisons

The mean percent correct identification scores shown for /ɹ/ and /l/ have been collapsed over the two confidence levels. Table 2 gives numerical summaries of scores for all words identified correctly.

Table 2

Average group scores on the /ɹ/ and /l/ identification task for "all correct" judgments (*SD* in parentheses) out of a possible score of eight

	NS	*EJ*	*IJ*
R in word list condition	7.9 (0.3)	6.9 (2.0)	7.0 (1.5)
L in word list condition	7.9 (0.2)	7.5 (0.6)	7.3 (1.2)
R in story condition	7.9 (0.1)	6.0 (2.5)	5.9 (2.3)
L in story condition	7.8 (0.3)	6.5 (1.4)	6.8 (1.2)

From Table 2 it can be seen that native speakers were judged to be performing equally well regardless of the sound (/ɹ/ and /l/) or the condition (story or word list). On the other hand, it appears that the two groups of Japanese speakers showed the same pattern: production was more accurate with /l/ than with /ɹ/, and scores were higher in the word list condition than in the story condition. Average scores, however, can be highly influenced by outliers. Boxplots (not given here for space considerations) showed that the data had outliers and was heavily skewed.

To determine whether there were significant differences between the groups of speakers, multiple comparisons were conducted among the three groups using 20% trimmed means and percentile bootstrapping. This approach is superior to using parametric statistics because the data are not normally distributed. Wilcox (2001, 2003) claims that one-way ANOVAs are not resistant to even small variations from non-normality or to heteroscedasticity (differences in variances). In contrast, means trimming is a robust answer to outliers in the data and bootstrapping the data leads to an empirical distribution that does not need to rely on normality (readers interested in these methods are referred to Wilcox, 2001). For all statistical tests the alpha level was set at .05. The statistical analysis was carried out using the R statistical software program[4]. Results will be reported as confidence intervals. The approach of using confidence intervals along with *p*-values is also generally preferred among statisticians, because confidence intervals provide not only information about whether the null hypothesis can be rejected or not, but also give an estimate of effect size (in how far away from zero the interval is) and confidence in results (the width of the confidence interval indicates the amount of sampling error) (Klein, 2004; Maxwell & Delaney, 2004).

Multiple comparisons[5] found no statistical differences between EJ and IJ for any of the conditions, but in each case found statistical differences between the EJ and NS and the IJ and NS (in the case of R in the word list condition for the EJ vs. NS the difference is marginally statistical). The confidence intervals and adjusted *p*-values are reported for each comparison in Table 3.

Quantitative analysis shows that the NS performed significantly better on all of the tests than either of the Japanese groups, but these groups did not significantly differ on those tests. If only the "definitely correct" identifications are examined, the results show the same trend, except that scores for NS range from 1–3% lower but scores for the Japanese groups are about 10% lower. Additionally, multiple comparisons showed the same trend as for the "all correct" data but even more strongly, with all *p*-values for the comparisons between the non-native and native groups being $p = .0000$. In this calculation the comparison between the EJ and NS for the R in word list condition was significant, with a confidence interval of (−3.19, −.81).

[4] The R program is free statistical software available from <*http://www.r-project.org*>. R is supported by a large community of statisticians, and is the free version of the S-PLUS software

[5] No omnibus main effects test was performed first. Wilcox (2003, p.415) notes that although omnibus tests are often undertaken before multiple comparisons, this effectively lowers the family-wise error rate, which may mask true differences among groups. Family-wise error rate was controlled by using the Benjamini-Hochberg adjustment for *p*-values (Benjamini & Hochberg, 1995).

Table 3

Mean difference between the Native Speaker, Experienced Japanese and Inexperienced Japanese groups ($n = 15$ in every group) in the four conditions for the word identification task

	EJ versus IJ	EJ versus NS	IJ versus NS
R in word list condition	(− 1.46, 1.58) p =.765	(− 1.93, .039) p =.07	(− 1.87,−.02) p =.024
L in word list condition	(−.67, .79) p =.65	(−.92,−.11) p =.000	(− 1.16,−.07) p =.01
R in story condition	(− 2.12, 2.76) p =.77	(− 3.40,−.15) p =.000)	(− 3.46,−.39 p =.000
L in story condition	(− 1.58, .88) p =.65	(− 2.28,−.28) p =.005	(− 1.67,−.25) p =.003

3.2.2 Effect of attention to task

The data were analyzed using both a parametric method and a bootstrapping method. A (2) Sound × (2) condition × (3) group repeated measures ANOVA for the "all correct" calculation method found that there was a statistical difference for Group, $F(2, 42) = 6.8$, p =.003, and Condition, $F(1, 42) = 29.5$, p <.0005, but not Sound (p =.13). There was also an interaction between Condition and Group, $F(2, 42) = 6.3$, p =.004. Multiple comparisons revealed that the EJ and IJ performed statistically differently (p <.001) on the two conditions while the NS did not ($p = 1.0$). A post hoc Tukey test for group revealed that the NE speakers were statistically different from the EJ and IJ speakers (p =.003 in both cases), but the EJ and IJ were not statistically significant (p =.991). Using 20% trimmed means and bootstrapping, NS were at ceiling and thus could not be compared, while none of the comparisons were statistical (p <.05) for either of the non-native groups. The reason both analyses are shown is to demonstrate that a researcher's choice of statistics can significantly affect their conclusions. The conclusion drawn here is that the non-native speakers tested do not perform statistically more accurately in producing /ɹ/ and /l/ when they are presumably paying less attention to form.

3.2.3 Judge's results

As was done in FTM (1995), a listener-based score was computed by averaging the judges' responses over the 15 speakers in each group. The 72 listener-based scores (2 sounds × 3 groups × 12 listeners) for "Definitely correct" answers were submitted to a repeated measures ANOVA with Group as the independent variable (between-subjects factors) and Sound (either /ɹ/ or /l/) as the dependent variable.[6] Significant

[6] After arcsine transformation, the scores for the dependent variables were approximately normally distributed, with R skewness = −.02, kurtosis = −.72, L skewness = −.29, kurtosis = −.54. A Levene's test for equality of variances found variances to be equal. The assumptions for ANOVA were thus upheld.

effects, with a large effect size, were found for group, $F(2, 33) = 27.7$, $p < .001$; partial eta-squared = .63, but not for sound ($p = .17$). There was a significant interaction between sound and group, $F(2, 33) = 3.5$, $p = .04$; partial eta-squared = .18. Tukey's post hoc tests for group revealed that NS performed differently from both EJ and IJ on both sounds ($p < .001$), while the Japanese speakers did not differ from one another ($p = .886$). These results show that when using a four-point scale the judges clearly distinguished between the native and Japanese groups.

Additionally, inter-rater reliability over all words was high, at 94.2% internal consistency. Reliability was computed using the Alpha reliability analysis, two-way mixed design, absolute agreement definition of intraclass correlation coefficients in the SPSS statistical program. This reliability percentage indicates the extent to which judges agreed on ratings of words.

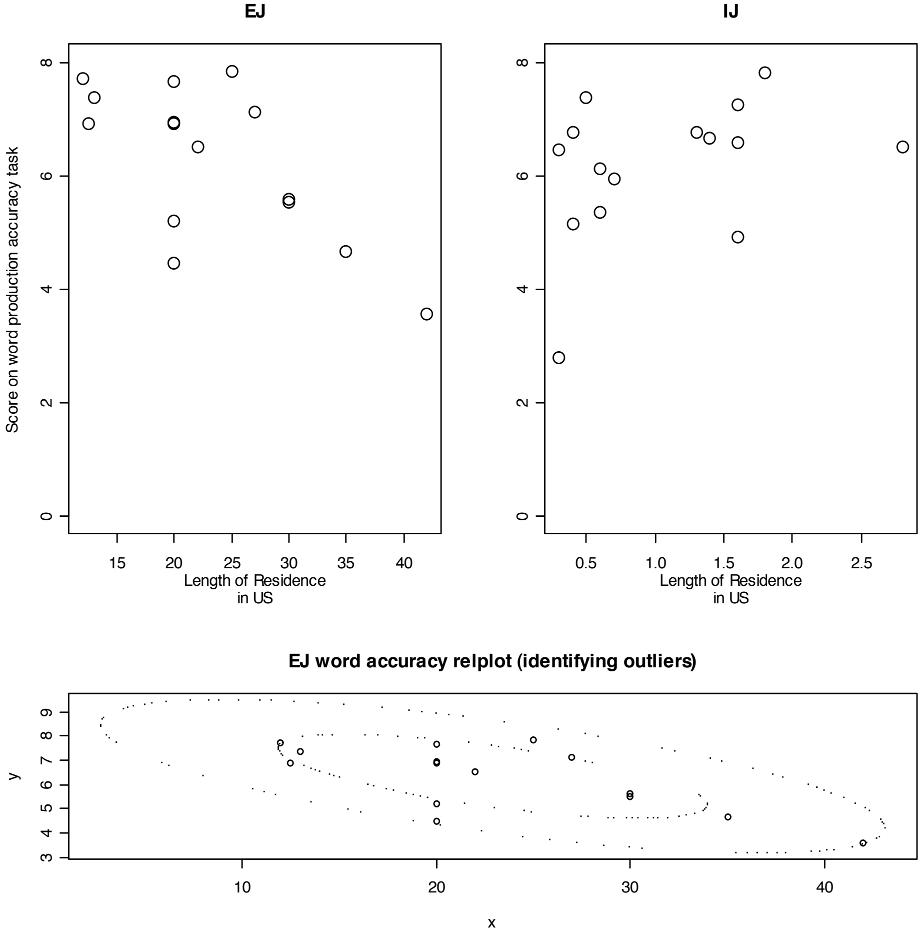
3.2.4 Correlational analyses

Contrary to the hypothesis that LOR can improve phonological accuracy, a negative relationship was found between LOR and a composite score for this task (the average of all 8 scores on this measure). Correlation was tested by using a percentile bootstrap method, which does not assume normality or homogeneity of variances (see Wilcox, 2003 for more information about this procedure). The correlation between LOR and scores on the task was not significant across all Japanese learners together ($r = -.06$, $p = .66$), but when the EJ and IJ were separated, the correlation for the IJ speakers was positive but not significant ($r = .37$, $p = .11$) while the correlation for EJ speakers became strong and marginally significant ($r = -.58$, $p = .06$). Figure 1 shows scores for both groups on a scatterplot. From the graph for the EJ it appears that a few outliers may be influencing the correlation to make it negative. A robust regression (using S-PLUS), which would reduce the influence of outliers on a small data set, found that the best fit line ($y = 8.8 - .11$ * LOR) was virtually identical to the least squares fit ($y = 9.1 - .12$ * LOR). This means that the correlational results are likely due to real effects and are not the influence of a few outliers. Additionally, a so-called relplot, which is a two-variable analog of the boxplot and is a method of detecting outliers (Wilcox, 2003), shows that none of the EJ participants is judged an outlier using this method either.

One might wonder why accuracy in producing words that begin in /ɹ/ and /l/ would decline with LOR. One logical possibility is that chronological age may be playing a role. After all, LOR clearly cannot increase without chronological age increasing, and in fact, these two factors are highly correlated for the EJ group ($r = .76$, $p < .0005$). The correlation between chronological age and scores on this task are marginally significant for the EJ ($r = -.37$, $p = .11$), and not as strong as for LOR. In order to see if some other factor might be simultaneously affecting the correlation with age but not LOR, partial correlations were performed separately, controlling for status of marriage to a native speaker, age of arrival in the U.S., and reported use of English. The partial correlation which resulted in the age and LOR correlations being closest in strength was the one involving Age Of Arrival (AOA). This makes sense, as for

Figure 1

Scatterplot of LOR values and scores on a production task involving /ɹ/ and /l/ for experienced Japanese participants (EJ) and inexperienced Japanese (IJ) ($n = 15$ for each group[8]; LOR measured in number of years). The relplot identifies as outliers any points which are outside of the second ellipse.



all but one of the EJ participants, chronological age was equal to AOA plus LOR.[7] With AOA factored out, the correlation between age and scores on this task was stronger and significant ($r = -.60$, $p = .023$, $df = 12$) and was quite close in strength to the bootstrapped correlation between LOR and scores on the task given above ($r = -.58$). On the other hand, a partial correlation of AOA with scores while controlling for age was positive, not as strong and had a higher p-value ($r = .51$, $p = .06$). An

[7] Chronological age was not equal to AOA + LOR for one participant (SB222). This participant lived in the U.S. at two separate times.

[8] It is difficult to see from the graph but there are five data points at LOR = 20

examination of the scatterplot between age and AOA shows that the chronologically older participants also tended to have higher ages of arrival (r =.46, p =.16). Although this correlation need not be positive, it is influenced to some extent by the fact that the minimum LOR was 12 years, so that the younger participants necessarily had younger ages of arrival. In short, the fact that AOA correlates positively with scores is still due to age effects, because older participants tended to have older AOA in this sample. A correlation with AOA and scores was not significant (r =.21, p =.42). Thus, the strong negative correlation with LOR can be thought of as due mainly to negative effects for actual chronological age, after removing some positive effects for the age of arrival in the U.S. (the older they are, the later they arrived in general, since LOR = Age − AOA).

In order to formalize these results using a different statistical method, a standard linear multiple regression with the factors of LOR, AOA, use of English, hours of pronunciation training in /ɹ/ and /l/ and marriage to a NS were regressed onto the composite scores dependent variable. For the EJ, this regression found a significant R^2 =.64, but the only significant variable was LOR, with p =.007 (Intercept β = 5.2, SE = 2.4; LOR β = −.12, SE =.03). Using the same factors but substituting age for LOR, a linear regression found a significant R^2 =.59, with two significant variables of age (p =.01) and AOA (p =.02) (Intercept β = 5.4, SE = 2.6; Age β = −.13, SE =.04; AOA β =.22, SE =.02). To restate the conclusion of the previous paragraph, LOR thus seems to be a proxy for at least chronological age plus AOA, plus some other variable which was not measured here, since the regression with age and AOA does not explain quite as much of the variance as LOR does.

3.3 *Discussion*

3.3.1 Comparison with FTM (1995)

While Flege, Takagi, and Mann (1995) found that experienced Japanese speakers were able to produce /ɹ/ and /l/ similarly to native English speakers, this replication study found that the EJ and IJ consistently were grouped together and performed distinctly from NS. As for the effect of length of residence on production success, the present study found a *negative* correlation between time spent in the U.S. and ability to produce intelligible /ɹ/ and /l/ among the EJ group.

The question as to why this replication did not find the same results as for FTM cannot be definitively answered. A failure to replicate suggests that the original findings cannot be generalized. Where no particular bias in the original study can be identified as a reason for the lack of replicability, caution must be advised regarding the conditions under which the findings will hold.

Given the caveat that no firm conclusions can be drawn as to why results in the present study differ from FTM, there are several evident differences between the learners in the two studies. In the FTM study, the EJ and IJ groups differed in their reported use of English (EJ = 5.7, IJ = 2.9 average score on three 7-point scales at home, work, and social situations). In fact, the participants in FTM in fact did not overlap in their usage: those with short LOR ranged from mean usage scores of 1.3 to 3.7, while those with long LOR ranged from 4.3 – 7.0. In the present study, participants happened to be more evenly matched with regard to use of English (EJ = 4.4, IJ = 3.8 on a 7-point scale). Another difference was that FTM's subjects used English more often than this study's EJ speakers (*M* = 4.4 on a seven-point scale in this study, *M* = 5.7 in FTM). Chronological age was also a factor that differed between the two studies. The EJ participants in this study were quite a bit older at time of testing (*M* = 53.5yrs.) and the IJ younger (*M* = 29yrs.) than FTM (1995)'s speakers (EJ = 43.7yrs., IJ = 35yrs.). Another profile difference is that of age of arrival. The participants in this study were evenly matched for mean age of arrival in the U.S. (EJ = 28.7, IJ = 27.1), despite the fact that this effect was unintended, while Flege et al.'s participants differed by about 10 years in mean arrival age (EJ = 23 yrs, IJ = 33.5 yrs). Given the fact that the negative correlation with LOR in this study can be analyzed as primarily the result of negative effects of chronological age, the differences in age may be the main reason for the discrepancies between this study and FTM.

3.3.2
Attention to task

Experiment 1 was designed to test learners in one condition where they would not be aware of the focus on /ɹ/ and /l/ and another where the focus would be quite clear. It was hypothesized that learners would produce /ɹ/ and /l/ more accurately in the word list condition than in the story condition. The results do not support this conclusion, however. The Japanese groups produced /ɹ/ and /l/ just as accurately in both conditions, and their overall accuracy was fairly high.

3.3.3
Nativelike learners

Although as a group the EJ and IJ were different from the NS group, some individual Japanese participants did quite well on the tasks. There were three EJ speakers and one IJ speaker who consistently scored within two *SD*s of the native speaker mean on /ɹ/ and /l/ in both conditions, judged for both number of overall correct judgments and for definite judgments only.

Table 4 shows the lower bound of the interval for NS that was two *SD*s below the mean for both /ɹ/ and /l/ in the two different conditions and two confidence ratings. The fact that some individuals were able to score within NS range on this task is not surprising in light of previous /ɹ/ and /l/ production studies which found "good" producers in the group (Goto, 1971; Sheldon & Strange, 1982), and other studies of exceptional learners' phonology (Bongaerts, 1999; Bongaerts et al., 1995; Bongaerts et al., 2000; Bongaerts, van Summeren, Planken, & Schils, 1997). However, I would argue, along with Long (1990) and DeKeyser and Larson-Hall (2005), that more than limited samples of speech, such as those used in this experiment, are needed in order to conclude that a learner has reached a native-like proficiency in phonology. For this reason, the sentence-length production judged in Experiment 3 was included.

Table 4

Percentage correct scores of nativelike Japanese speakers on the R / L identification task

	Story, All Correct	*List, All Correct*	*Story, Def. Correct*	*List, Def. Correct*
NS (lower boundary of 2 *SD* from mean)	94.11	94.11	85.65	91.46
RI231 (IJ)	100	98.5	95.3	96.9
SB218 (EJ)	100	98.5	95.3	98.5
SB222 (EJ)	98.8	97.5	95	93.8
YM212 (EJ)	95.3	98.5	93.8	95.3

4 Experiment 2: The Word-Accent Rating Task

FTM (1995) found that the Japanese speakers with a long LOR produced /ɹ/ and /l/ tokens that were identified as accurately as those of NS, but they stated that high intelligibility did not always equate to "native-like" speech. Thus, they also tested word productions for their degree of foreign accent. They found that although the EJ's scores were not in the NS range, they were judged to be significantly better than the IJ. In this experiment groups were tested on the amount of *accent* they were perceived to have on words beginning in /ɹ/ and /l/. This experiment differs from Experiment 1, in which the *accuracy* with which words were produced was tested.

4.1 *Procedure*

A subset of 10 of the 16 words used in the R / L identification task were used in this test: *reek-leak, room-loom, rate-late, road-load,* and *rock-lock*. This sampled /ɹ/ and /l/ before high, mid, and low vowels. Only the words produced in the word list condition were used, but in this case they were remixed into a CV form. Following the procedure described in FTM (1995), the original word was cut to the CV form in order to eliminate any accent judging bias that might result from the later part of the word.

The judges in the word accent-rating task were the same as for the R / L identification task. The utterances were each judged twice. Thus each judge heard (10 words × 2 times × 15 speakers) = 300 words. The experiment was conducted by using a computer program, so all items were randomly presented, thus eliminating any ordering effects. Judges were told to rate the CV utterances on a scale from 1 – 7, where 1 was "strong foreign accent" and 7 was "no foreign accent." Judges were urged to make use of the entire scale. A box containing 'R' or 'L' was also visible on the computer screen, so that the judge would be aware of the identity of the intended production. The judges completed a trial use of the computer program with 12 CV utterances before they began the actual program.

4.2
Results

The results reported by mean accent ratings are found in Table 5. Overall inter-rater reliability was high at 88.8%. It can be seen that the native speakers scored more highly than the Japanese groups. Also, the IJ scored slightly higher than the EJ group. There does not seem to be much difference in accent across the /ɹ/ and /l/ words. Boxplots showed the data was not highly skewed although there were outliers in both the learner groups.

Table 5

Mean scores on word-accent rating task (*SD*s given in parentheses)

Rating (7 highest)	*NS*	*EJ*	*IJ*
/ɹ/	6.08 (.16)	3.79 (.22)	4.12 (.20)
/l/	5.84 (.16)	3.72 (.21)	4.12 (.28)

The data were submitted to a bootstrapped two-way ANOVA (Wilcox, 2003). This method allows for non-normality and heteroscedastic variances (not all the variances of the groups need to be equal). There were significant main effects for Group ($p < .0001$) but not Sound ($p = .37$), nor was the interaction between Sound and Group significant. Post hoc bootstrapped multiple comparisons found that, as in the previous experiment, the NS speakers were significantly different from the EJ and IJ speakers in both /ɹ/ and /l/ (/ɹ/: EJ versus NS (1.6, 2.9), $p = .000$; IJ versus NS (1.4, 2.6), $p = .000$; /l/: EJ versus NS (1.5, 2.9), $p = .000$; IJ versus NS (.8, 2.7), $p = .000$, but there was no significant difference between EJ and IJ speakers (/ɹ/: EJ versus IJ (−.9, .4), $p = .64$; /l/: EJ versus IJ (−1.4, .4), $p = .55$ (all *p*-values adjusted for multiple comparisons).

4.2.1
Correlational analyses

A marginally negative relationship was found between LOR and a composite score for word accent across all Japanese speakers of English (percentile bootstrapped correlation $r = -.35$, $p = .06$). Separating out EJ and IJ, however, the correlation for the IJ speakers became positive and significant ($r = .51$, $p = .04$) while the correlation for EJ speakers became strongly and highly significantly negative ($r = -.81$, $p < .0005$). As in Experiment 1, a relplot found no outliers among the EJ.

In looking again at the question of age, the correlation for EJ between overall score on this task and age was also negative and significant ($r = -.66$, $p = .003$), although not as strong as the correlation with LOR and scores. The partial correlation of age and scores while controlling for AOA is much stronger ($r = -.83$, $p < .0005$, $df = 12$). In fact, this is slightly stronger than the correlation with LOR and scores ($r = -.81$). The partial correlation between AOA and scores among the EJ, controlling for age, is also significant but positive and not as strong ($r = .71$, $p = .004$). A correlation between

AOA and scores for the EJ is not significant (r=.59, p=.15). The correlation between age and scores for IJ is not significant (r=−.1, p=.72).

4.3 *Discussion*

4.3.1 Comparison with FTM (1995)

Group distribution for the accent ratings again differed from FTM (1995), but was consistent with the findings in the previous task: NS scored consistently better than either the EJ or IJ, but these two latter groups were not significantly different from one another. FTM (1995) found that the NS and EJ were significantly different from each other in the accent rating task, but also that the EJ were significantly better than the IJ in the accent-ratings task. Although FTM (1995) argued that results from the identification task showed that LOR could lead Japanese speakers to achieve the same results as NS, the ratings task in that study showed that even very good speakers could not achieve as high ratings for accent as NS, even on very short segments of speech.

4.3.2 Correlational analyses

In contrast to the results for word accuracy (which found no correlation with LOR), for word accent there is a significant effect of LOR among the IJ. The effect is positive, meaning that the longer the inexperienced speakers had lived in the U.S., the better their accent on /ɹ/ or /l/ was judged to be. This effect is presumed to be due to the continuing input the learners received, as age did not correlate with scores.

On the other hand, for the EJ group the strong negative correlation between word accent scores and LOR can be seen as resulting mostly from chronological age, again with some positive effects for later age of arrival subtracted out.

4.3.3 Nativelike learners

Although as a group accent ratings for this task were lower for EJ and IJ speakers than for NS speakers, there was one IJ subject who fell within two *SD*s of the NS mean on *both* the /ɹ/ (lower bound: 4.86) and /l/ (lower bound: 4.63) words. This speaker (NN213) was not one of the Japanese speakers who fell within NS norms for the R/L identification task, however. There were four other EJ who fell within the NS bounds on either /ɹ/ or /l/ (2 for /ɹ/ and 2 for /l/) but none had done as well as NS in Experiment 1. There were also four IJ who fell within the NS bounds on /l/, but only one (RI213) had done as well as NS in Experiment 1. This is an interesting finding as it shows individual differences among the Japanese in learning to produce either /ɹ/ or /l/ well.

5 Experiment 3: The Sentence Rating Task

The purpose of this experiment was to see whether NS and EJ would be judged differently in a more global speaking condition, given the assumption that the EJ might do as well as NS in the previous tasks. In this experiment speakers were rated for global foreign accent on a sentence-length utterance taken from the same story task as used in Experiment 1. Global foreign accent ratings have often been used in testing for differences between younger and older language learners (Bongaerts, Mennen, & van der Slik, 2000; Flege, Yeni-Komshian & Liu, 1999; Thompson, 1991). Because Flege, Takagi and Mann (1995) reported that their EJ were close to NS norms,this experiment was included as a way to see whether the EJ participants could come close to NS norms on a longer speech sample.

5.1
Method

The same speakers and judges were used as in the previous two experiments. In this experiment, every judge rated all 45 of the speakers. The judges were asked to use the same 1–7 accent rating scale that was used in the Experiment 2, where a score of one meant "strong foreign accent" and seven meant "no foreign accent."

5.1.1
Stimulus preparation

The sentence for task three was taken from the story reading condition. The sentence was, "Before she knew it she fell asleep, and when she woke up it was very late." This sentence was chosen because no speakers had difficulties saying the entire sentence without any mistakes. It was located in the middle of a paragraph that was itself in the middle of the story (see Appendix 2). No attempt was made to control for speaking rate although very long pauses were deleted inside the sentence. Munro and Derwing (2001) found that speaking rate is a confounding factor in judging the accentedness of sentences such as this, accounting for 6–15% of the variance in listeners' scores.

5.2
Results

Scores on the sentence rating task were averaged for each speaker across all 12 judges. The mean score for the NS was 6.79 (*SD*=.17), for the EJ was 3.18 (1.16) and for the IJ was 3.29 (1.22). Multiple comparisons using 20% trimmed means and bootstrapping were performed for the three groups. The NS were found to differ significantly from the EJ (confidence interval (2.85, 4.45), $p<.005$) and the IJ (confidence interval (2.59, 4.46), $p<.005$). However, the EJ and IJ were not significantly different from one another (confidence interval (−1.31, 1.06), $p=.72$) (p-values are adjusted for multiple comparisons).

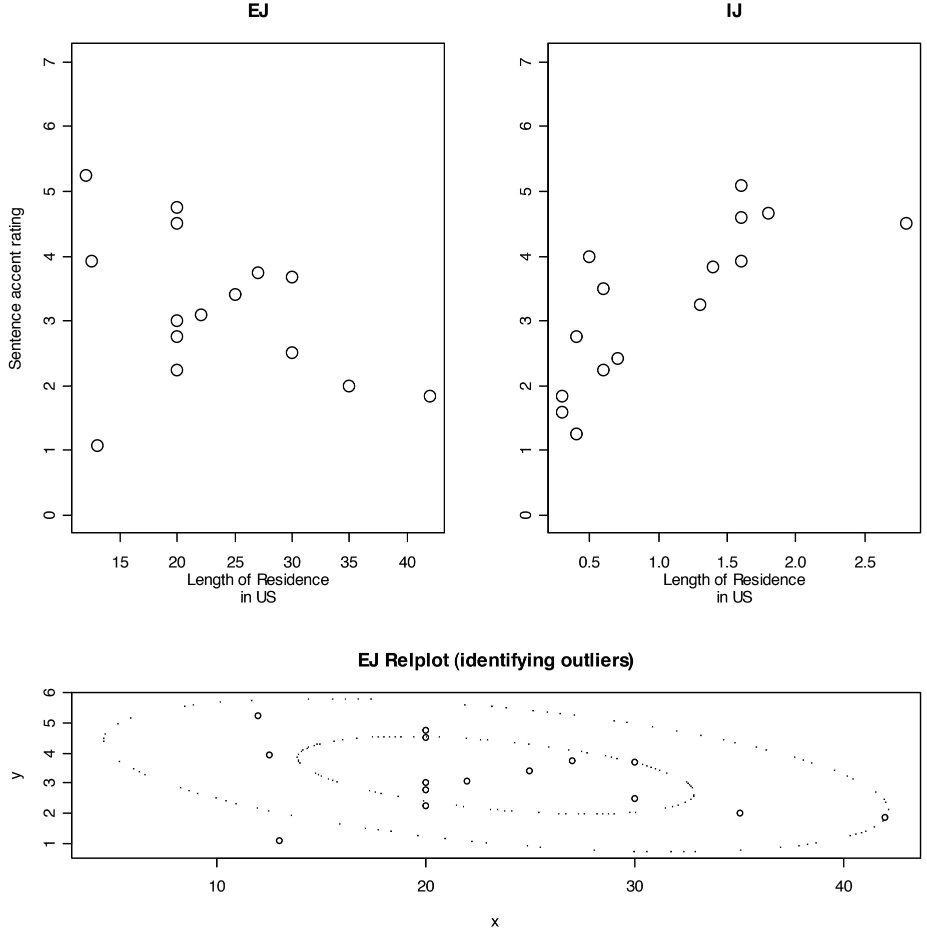
This finding continues the trend found in the previous two experiments.

5.2.1
Correlational analyses

When the non-native groups are examined separately, there is a large and strong positive correlation between LOR and sentence accent for the IJ ($r=.85, p<.0005$). For the EJ, the correlation is not significant ($r=-.36, p=.29$). However, it appears there is an outlier in the EJ group. Figure 2 shows scatterplots for the EJ and IJ groups and a relplot for the EJ. The relplot shows that the EJ participant whose LOR was 13 years and whose score on the sentence accent task was 1.08 lies outside of the outer ellipsis, and is thus an outlier. Without this participant, the correlation for the EJ becomes marginally significant ($r = -.58$, $p = .06$).

Figure 2

Scatterplot of LOR values and rated sentence accent for words beginning with /ɹ/ and /l/ for experienced Japanese participants (EJ) and inexperienced Japanese (IJ) ($n = 15$ for each group; LOR measured in number of years). A relplot for the EJ show an outlier at $y = 1.08$



The relationship between age and sentence accent ratings was not significant for the IJ ($p = .17$), but was marginally significant for the EJ ($r = -.47$, $p = .07$). A partial correlation controlling for AOA revealed that the correlation became significant and practically the same strength as the LOR correlation ($r = -.57$, $p = .04$, $df = 11$). On the other hand, the partial correlation of AOA and sentence accent scores controlling for age was nonsignificant ($r = .33$, $p = .28$, $df = 11$). There was no correlation between AOA and scores for the EJ ($r = -.01$, $p = .93$).

5.3 *Discussion*

5.3.1 Comparison with FTM (1995)

As in the previous two experiments, the sentence accent rating showed that non-native Japanese learners of English scored differently from NS. It also showed that extensive residence in the U.S. did not lead the EJ group to higher scores than the IJ group. LOR contributed differently to the success of the two groups; for the IJ, the longer they lived in the U.S., the better their sentence accent ratings became. This was a very strong correlation, accounting for 72% of the variance in the data. This agrees with the findings of Piske et al. (2001) that practice and input, obtained in the target country, may help make a difference across the entire phonology in initial stages of residence. Age, however, had nothing to do with success for the IJ.

On the other hand, LOR tended to negatively affect scores for the EJ; in general, the longer they lived in the U.S., the worse their scores became (although one participant with LOR 13 years, labeled an outlier, was worst of all). This was shown to be due in large part to increasing age. Experiment 3 finds no support for the assertion that long-term LOR and the implication of continued large amounts of input can help these speakers substantially improve their phonology across sentences.

5.3.2 Nativelike learners

In this task there were no Japanese speakers who scored within the range of NS. The lower boundary of the native speaker 95% confidence interval was 6.6, and the highest mean score received by a Japanese speaker was 5.25 by one IJ speaker (SB222).

6 General discussion

The first research question was whether LOR would correlate positively with production accuracy and accentedness on the phonological measures tested here. Correlational analyses in all three experiments identified a significant or near-significant negative correlation between LOR and scores for the Japanese users of English with long-term residence. This surprising finding indicates that more research with older L2 learners is warranted if we want to understand the entire spectrum of second language learning in adults.

It is plausible to suppose that the decline in scores with longer LOR was due to chronological age. The present study documents a clear negative correlation between phonological scores in all three experiments and chronological age, when age of arrival is controlled for. This is a curious result, given that although studies of elderly people have found clear declines in speech *perception*, mostly related to hearing loss (van Rooij & Plomp, 1990, 1992 for L1), no previous studies have documented a clear loss of *production* accuracy for older participants, whether in the L1 or the L2. The available studies point to no clear conclusions. For L1, Sweeting and Baken (1982) found no significant difference in average voice onset time measures between younger and older native speakers of French. However, Ryalls, Cliche, Fortier-Blanc, and

Prud'Hommeaux (1997) found that younger French speakers ($M = 24.2$) produced larger differences in voice onset time than older speakers ($M = 67.1$). For L2, Scott (1994) compared younger and older English-Spanish bilinguals' performance on a variety of tasks including a production task where participants repeated words, phrases and sentences in Spanish for a total of 30 items. The productions were transcribed by the author, who offered the undetailed conclusion that "younger bilinguals produced more phonemes and performed with much greater accuracy then the older bilinguals" (1994, p. 272).

It is possible that the aging process itself may affect production. There is evidence that with age, adults become less able to control executive functions such as controlling attention and ignoring irrelevant stimuli, unless responses are highly automatized (Craik & Jacoby, 1996; Zacks, Hasher, & Li, 2000). Bialystok, Craik, Klein and Viswanathan (2004) have found that although balanced bilingualism confers some benefits in processing control that counteract the cognitive decline of age, both bilinguals and monolinguals age 60 and older do show declines in reaction times and inhibitory control in psycholinguistic experiments. It is possible that L2 users become less accurate in production as they age, due to less central processing control and less automaticity in their production of difficult sounds. Another possibility is that hearing loss could be related to declines in production as well as in perception accuracy. Yet another possibility is that the participants may differ with regards to psychological factors such as identity or attitude. The oldest participants in this group lived in a very different historical moment (one closer to World War II) than the younger participants, and this may have affected their attitudes toward learning English. Clearly this study does little more than point up the need for more research in this area.

For the Japanese learners of English with short-term residency, on the other hand, LOR did correlate positively with word and sentence accent, while age did not correlate with these measures. This study provides solid evidence that an initial one to two years of immersion in the target-language country can lead to gains in nativelike accent for words beginning with /ɹ/ and /l/ and also for entire sentences.

The second research question was whether Japanese participants with extensive residence would be judged to produce /ɹ/ and /l/ as accurately as NS of English, and whether the accentedness of their words beginning in /ɹ/ and /l/ as well as entire sentences would be judged to be within the NS range. Contrary to the results found by Flege, Takagi and Mann (1995), this study failed to find evidence that long-term residence in a target-language country could lead to accuracy that is equal to that of native speakers. In addition, again contrary to FTM, in group analyses, this study failed to find that the long-term residents were any better or more accurate producers of /ɹ/ and /l/ or of entire sentences than Japanese learners of English who had lived in the U.S. for only one year on average.

This may appear to contradict studies like Flege and Fletcher (1992) and Flege, Bohn and Jang (1997), which found that learners with extended residence in the U.S. were more accurate phonologically than learners with very short residences. However, findings from this study suggest that participants may receive substantial improvement in their phonology, especially for entire sentences, over the course of at least two years. In the studies by Flege and colleagues the inexperienced speaker groups had average lengths of residence of less than a year. Thus, perhaps the superiority of more experienced groups may have been due more to the fact that the inexperienced groups had not yet reached the peak of their potential rather than to any significant long-term gains.

In sum, extensive LOR and its attendant implication of extensive input did not prove to be a good predictor of performance for the participants in this study. It is possible that more sensitive measures of input could be more predictive of outcomes, although this is unlikely given practical memory constraints in remembering how much input was received long ago.

The third research question was whether Japanese users of English would produce /ɹ/ and /l/ more accurately on a word list, which presumably would direct their attention to the sounds, than in a story-reading task where they had to focus on reading at a fluent rate. It was found that the Japanese actually scored numerically higher on the story-reading condition, but that there was no statistical difference between scores in the two conditions. The conclusion is that learners perform equally well in both conditions, and that attention to form, at least for these Japanese learners with the /ɹ/ and /l/ sounds, does not enhance performance.

6.1
Questions for future study

Future research should be directed at developing more detailed ways of measuring levels of input. Questions like those found in Moyer (2004), which asked about typical amount of time spent watching TV and listening to, reading in, and speaking English each week, can be a guide. It would be very interesting to have more studies, especially longitudinal studies which retested the same learners every year or six months, that would pinpoint exactly how much input in an immersion environment leads to beneficial effects in the short-term. Lastly, much research remains to be done with older users of a second language to test whether the results found in this study are generalizable to other populations, and whether language production accuracy declines in the L1, L2, or both. Such research should include more measures of attitudes and individual differences than were found in this study or in FTM. These will probably have to be cross-sectional studies, since testing learners, say, once a decade as they age is not a feasible research design.

manuscript received: 05. 18, 2005
manuscript accepted: 02. 28, 2006

References

BENJAMINI, Y., & HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, B, **57**, 289–300.

BIALYSTOK, E., CRAIK, F. I. M., KLEIN, R., & VISWANATHAN, M. (2004). Bilingualism, aging, and cognitive control: Evidence from the Simon task. *Psychology and Aging*, **19**(2), 290–303.

BONGAERTS, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late L2 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**(2), 298–308.

BONGAERTS, T., PLANKEN, B., & SCHILS, E. (1995). Can late learners 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). Bristol, PA: Multilingual Matters.

BONGAERTS, T., van SUMMEREN, C., PLANKEN, B., & SCHILS, E. (1997). Age and ultimate attainment in the pronunciation of a foreign language. *SSLA*, **19**, 447–465.

CHAUDRON, C. (1983). Research on metalinguistic judgments: A review of theory, methods, and results. *Language Learning*, **33**(3), 343–377.

CRAIK, F. I. M., & JACOBY, L. L. (1996). Aging and memory: Implications for skilled performance. In W. A. Rogers, A. D. Fisk & N. Walker (Eds.), *Aging and skilled performance: Advanced in theory and applications* (pp. 113–137). Mahwah, NJ: Erlbaum.

DEKEYSER, R. M., & LARSON-HALL, J. (2005). What does the Critical Period really mean? In J. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches*. New York: Oxford University Press.

DICKERSON, L., & DICKERSON, W. (1977). Interlanguage phonology: Current research and future directions. The notions of simplification, interlanguages, and pidgins, and their relation to second language pedagogy, *Actes du 5eme Colloque de Linguistique Appliquee de Neuchatel* (pp. 18–30). Geneva: Droz.

DŐRNYEI, Z. (2005). *The psychology of the language learner*. Mahwah, NJ: Lawrence Erlbaum Associates.

FLEGE, J. E. (1987). A critical period for learning to pronounce foreign languages? *Applied Linguistics*, **8**(2), 162–177.

FLEGE, J. E. (1999). Age of learning and second language speech. In D. Birdsong (Ed.), *Second language acquisition and the Critical Period hypothesis* (pp. 101–131). Mahwah, NJ: Erlbaum.

FLEGE, J. E., BOHN, O. S., & JANG, S. (1997). The production and perception of English vowels. *Journal of Phonetics*, **25**, 437–470.

FLEGE, J. E., & FLETCHER, K. L. (1992). Talker and listener effects on degree of perceived accent. *Journal of the Acoustical Society of America*, **91**(1), 370–389.

FLEGE, J. E., & LIU, S. (2001). The effect of experience on adults' acquisition of a second language. *Studies in Second Language Acquisition*, **23**, 527–552.

FLEGE, J. E., & MacKAY, I. R. A. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition*, **26**(1), 1–34.

FLEGE, J. E., MacKAY, I. R. A., & PISKE, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics*, **23**, 567–598.

FLEGE, J. E., MUNRO, M. J., & MacKAY, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, **97**(5), 3125–3138.

FLEGE, J. E., TAKAGI, N., & MANN, V. A. (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. *Language and Speech*, **38**(1), 25–55.

FLEGE, J. E., YENI-KOMSHIAN, G., & LIU, S. (1999). Age constraints on second-language acquisition. *Journal of Memory and Language*, **41**, 78–104.

GOTO, H. (1971). Auditory perception by normal Japanese adults of the sounds 'l' and 'r'. *Neuropsychologia*, **9**, 317–323.

GUION, S. G., FLEGE, J. E., & LOFTIN, J. D. (2000). The effect of L1 use on pronunciation in Quichua-Spanish bilinguals. *Journal of Phonetics*, **28**(1), 27–42.

HYLTENSTAM, K., & ABRAHAMSSON, N. (2003). Maturational constraints in SLA. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 539–588). Malden, MA: Blackwell.

JIA, G. X., & AARONSON, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics*, **24**, 131–161.

JIA, G. X., AARONSON, D., & WU, Y. (2002). Long-term language attainment of bilingual immigrants: Predictive variables and language group differences. *Applied Psycholinguistics*, **23**, 599–621.

KLEIN, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

LONG, M. H. (1990). Maturational constraints on language development. *SSLA*, **12**(3), 251–285.

MAJOR, R. C. (1994). Current trends in interlanguage phonology. In M. Yavas (Ed.), *First and second language phonology* (pp. 181–203). San Diego: Singular.

MAXWELL, S. E., & DELANEY, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: LEA.

McMANUS, C., & PLEINES, C. (2005). *Personality, emotional intelligence, motivations and barriers in second language learning.* Paper presented at the Second Language Research Forum, New York.

MOYER, A. (2004). *Age, accent and experience in second language acquisition*. Clevedon: Multilingual Matters.

MUNRO, M. J., & DERWING, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, **23**, 451–468.

OYAMA, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, **5**, 261–283.

PISKE, T., FLEGE, J. E., MacKAY, I. R. A., & MEADOR, D. (2002). The production of English vowels by fluent early and late Italian-English bilinguals. *Phonetica*, **59**, 49–71.

PISKE, T., MacKAY, I. R. A., & FLEGE, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, **29**, 191–215.

RYALLS, J., CLICHE, A., FORTIER-BLANC, I. C., & PRUD'HOMMEAUX, A. (1997). Voice-onset time in younger and older French-speaking Canadians. *Clinical Linguistics and Phonetics*, **11**(3), 205–212.

SATO, C. (1985). Task variation in interlanguage phonology. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 181–196). Rowley, MA: Newbury House.

SCOTT, M. L. (1994). Auditory memory and perception in younger and older adult second language learners. *Studies in Second Language Acquisition*, **16**, 263–281.

SHELDON, A., & STRANGE, W. (1982). The acquisition of /ɹ/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, **3**, 243–261.

SKEHAN, P., & FOSTER, P. (2001). Cognition and tasks. In M. H. Long & J. C. Richards (Eds.), *Cognition and Second Language Instruction* (pp. 183–205). New York: Cambridge University Press.

SWEETING, P., & BAKEN, R. (1982). Voice onset time in a normal-aged population. *Journal of Speech and Hearing Research*, **25**, 129–134.

THOMPSON, I. (1991). Foreign accents revisited: The English pronunciation of Russian immigrants. *Language Learning*, **41**(2), 177–204.

Van ROOIJ, J. C. G. M., & PLOMP, R. (1990). Auditive and cognitive factors in speech perception by elderly listeners. II: Multivariate analyses. *Journal of the Acoustical Society of America*, **88**, 2611–2624.

Van ROOIJ, J. C. G. M., & PLOMP, R. (1992). Auditive and cognitive factors in speech perception by elderly listeners. III. Additional data and final discussion. *Journal of the Acoustical Society of America*, **91**(2), 1028–1033.

WILCOX, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer.

WILCOX, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Elsevier Science.

YENI-KOMSHIAN, G., FLEGE, J. E., & LIU, S. (2000). Pronunciation proficiency in the first and second languages of Korean-English bilinguals. *Bilingualism: Language and Cognition*, **3**(2), 131–149.

ZACKS, R. T., HASHER, L., & LI, K. Z. H. (2000). Human memory. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (2nd ed., pp.293–357). Mahwah, NJ: Erlbaum.

Appendix 1: Individual data

Table 6

Main variables for Japanese subjects with short LOR (IJ)

ID	*AOA*	*Age at test*	*LOR*	*Use of English (out of 21)*	*Married to NS?*	*Pron. instr. hrs*	*Exp. 1 (ID)*	*Exp. 2 (accent) (7 best)*	*Exp. 3 (Global pron.) (7 best)*
CK203	18	20	1.6	12	No	.17	61	4.58	5.08
HN247	33	33	.4	19	Yes	5	64	3.99	1.25
KN206	26	26	.4	12	No	0	84	4.05	2.75
KO207	27	28	1.6	8	No	84	82	4.55	3.92
MI246	19*	36	1.3	9	No	0	84	3.98	3.25
MK202	31	31	.5	15	No	12	92	4.74	4.00
MM209	31	31	.3	8	No	1	81	3.30	1.58
NK216	31	31	.6	20	No	.25	77	4.60	2.25
NN213	17*	22	1.6	19	No	1	91	5.65	4.58
RI231	30	32	1.8	11	No	12	98	5.18	4.67
SK208	31	31	.6	9	No	0	67	3.14	3.50
ST211	31	31	.7	11	No	1	74	4.23	2.42
TM210	28	28	.3	4	No	.08	35	2.24	1.83
YI204	25	26	1.4	10	No	.5	83	3.34	3.83
YY201	28	30	2.8	5	No	0	81	4.23	4.50

*These participants lived in English-speaking countries for the length of residence listed, but did so at two separate times.

Table 7

Main variables for Japanese subjects with long LOR (EJ)

ID	*AOA*	*Age at test*	*LOR*	*Use of English (out of 21)*	*Married to NS?*	*Pron. instr. hrs*	*Exp. 1 (ID) (8pt)*	*Exp. 2 (accent) (7 pts)*	*Exp. 3 (Global pron.) (7 pts)*
EG227	25	46	20	16	Yes	1	6.91	4.44	3.00
FS228	37	50	13	10	No	.5	738	4.76	1.08
HD230	26	38	12.5	7	Yes	4	6.91	3.94	3.92
JF233	22	55	35	12	Yes	.17	4.66	2.61	2.00
KL220	22	53	30	21	Yes	.5	5.59	3.10	3.67
KM214	37	59	22	10	Yes	0	6.50	3.50	3.08
KM221	35	65	30	6	No	1	5.53	3.21	2.50
MM217	38	65	27	5	No	20	7.13	3.78	3.75
MS226	25	45	20	6	No	0	5.21	4.47	2.25
SB218	29	55	25	21	Yes	1	7.84	3.56	3.42
SB222*	25	45	12	15	Yes	0	7.70	4.25	5.25
SC215	28	48	20	15	No	.25	6.94	4.69	4.75
SN205	33	53	20	9	No	24	4.46	3.62	2.75
TH219	24	66	42	21	Yes	0	3.56	2.46	1.83
YM212	25	45	20	17	No	.17	7.66	3.95	4.50

* These participants lived in the U.S. for the length of residence listed, but did so at two separate times.

Appendix 2: Story-Reading Task

Jane was a woman who had a lot of bad luck. One time when she had just bought a new dress, a rip appeared out of nowhere! Perhaps it happened because of the ride in Jimmy's truck, or maybe the fence she climbed over. Whichever one, she couldn't tell her husband Bob she had torn her dress. So she lied.

Another time she had to go to the hospital because she cut herself opening a can. The rim was sharp when she pulled the lip of the can up, and soon red blood was running down her right hand. She asked Bob to take a look at it, and he thought they had better go to the emergency room. On their way out to the car, Jane fell on a rake. She lay there a minute until Bob helped her up. She was covered with dirt and felt very low. The doctor said it was only a small injury and gave her a bandaid. She felt very foolish!

Jane lived in a very old apartment, and one day she called the manager to get him to fix the leak in the bathroom faucet. When a repairman came later that day, Jane showed him to the room with the leak. The man undid his ruck and got his tools out, and asked Jane to shine his light at the pipes so he could see.

"There's a lot of dirt in these pipes," he announced. "I can't understand why it doesn't reek!"

"Well, we don't let things rot in there purposely," she declared, slightly annoyed.

While the man cleaned out the pipes, Jane quietly sat in the living room, because she had to read her chemistry book. Before she knew it, she fell asleep and when she woke up, it was very late. It was dark, and she wondered if the plumber were still there, or if he had hit the road without even saying goodbye. As she got up, she noticed that several things were gone — the radio, TV and some pearls she had found in the Tampa Reef on her Florida vacation! She couldn't believe that the repairman had taken her loot! She called the manager but he said he had not called a repairman yet. Jane was very upset at being robbed and made Bob install a new lock in the door.

Another bad thing happened when Bob and Jane took a vacation to Lancaster, PA. They piled their tent onto the rack of their car (they didn't lack anything) and set out for an adventure. They saw museums of how people made things long ago, like when they sewed their clothes on a loom. They saw a row of horses plowing a field. Near a pretty lake they got out of the car to watch the squirrels run around the trees. The squirrels ran around on the ground, over the roots of the trees, and then scrambled up the trees to walk out on the thin limbs to gather leaves and nuts. Before Jane knew it, two squirrels were throwing nuts at her head! Although the nuts were small, it hurt and then she fell on a rock and sprained her ankle!

As the last ray of sun left the horizon, Jane and Bob driving to the hospital once more, with Bob driving at a fast rate. They were surprised when a policeman pulled them over for speeding! Bob explained their situation, and had the policeman lead them to the hospital. Soon the doctor fixed her cut and told her to keep her load off her ankle.

Although Bob and Jane had to spend the rest of their vacation in the hospital, they had a lot of time to play their favorite game — chess! Bob had a winning way with his rook, but Jane always beat him with her queen. And so end the unlucky tales of Jane.