

Second Language Fluency: Judgments on Different Tasks

Tracey M. Derwing and Marian J. Rossiter
University of Alberta

Murray J. Munro
Simon Fraser University

Ron I. Thomson
University of Alberta

In this study we determined whether untrained raters' assessments of fluency in low-proficiency second language speech were related to temporal measures and whether they varied across tasks. We collected speech samples from 20 beginner Mandarin learners of English on picture description, monologue, and dialogue tasks. Temporal measures were made on each sample. Twenty-eight untrained judges rated fluency, comprehensibility, and accentedness. Three trained raters also judged samples for "goodness of prosody." The rating data paralleled the

Tracey Derwing and Marian Rossiter, Department of Educational Psychology; Murray Munro, Department of Linguistics; Ron Thomson, Department of Linguistics.

The authors acknowledge the students and staff at NorQuest College; without their willing participation, this study would not have been possible. We also thank Cynthia Lambertson, Amy Stancliff, Adele Tan, and Amy Meckelborg for their assistance and three anonymous reviewers for their helpful comments. Finally, we are grateful to the Social Sciences and Humanities Research Council of Canada for funding this study.

An earlier version of this article was presented at the American Association of Applied Linguistics on March 24, 2003, in Arlington, Virginia.

Correspondence concerning this article should be addressed to Tracey Derwing, Department of Educational Psychology, University of Alberta, Edmonton, AB T6G 2G5, Canada. Internet: tracey.derwing@ualberta.ca

speech measurements: speakers' performance on the monologue and dialogue tasks was significantly better than on the narratives; however, listeners' judgments of goodness of prosody did not vary across tasks. Comprehensibility and fluency ratings were highly correlated; fluency was more strongly related to comprehensibility than to accentedness.

As a number of researchers have observed (Chambers, 1997; Guillot, 1999; Schmidt, 1992; Wood, 2001), not only is the term *fluency* difficult to define, but it has a wide range of definitions associated with it. For the purposes of this study, we will consider second language (L2) fluency to be an automatic procedural skill (Schmidt, 1992) on the part of the speaker and a perceptual phenomenon in the listener. As Segalowitz (2004) has pointed out, using Levelt's (1989) speech production model, an L2 speaker's fluency has its beginnings in the formulator, in which lexical access, phonological short-term memory, and control of attention all influence the eventual productions of the articulator. In the current study we are interested in examining the product of the formulator and articulator and its impact on the listener. In particular, we are concerned primarily with temporal variables that reflect the degree of effort L2 learners exert in order to produce utterances under a range of conditions. We are also interested in the extent to which those same temporal factors have an effect on listeners.

For both theoretical and practical reasons, research should be carried out to establish the factors that contribute to perceptions of L2 fluency, the reliability of judgments of fluency, and the extent to which untrained listeners' judgments correspond to those of trained listeners. L2 learners' fluency is often assessed in high-stakes tests (e.g., the Test of Spoken English) that have a tangible effect on learners' futures. Oral tests are regularly a determining factor in university admission and employment decisions.

In recent years there has been a growing interest in L2 fluency from the standpoint of the speaker, as is evident in the

research of Foster and Skehan (1996), Lennon (1990, 2000), Skehan and Foster (1999), Towell, Hawkins, and Bazergui (1996), Riggenbach (1991, 2000), Wennerstrom (2000), and others. Studies have focused on a variety of factors that affect fluency, such as task type (e.g., Bygate, 1996; Ejzenberg, 1992; Skehan & Foster, 1999), preplanning (e.g., Crookes, 1989; Foster & Skehan, 1996; Mehnert, 1998; Ortega, 1999; Wigglesworth, 1997), online planning (Yuan & Ellis, 2003) and the planning process (Ortega, 1999), time spent in an L2 environment (e.g., Freed, 1995; Lennon, 1990; Towell et al., 1996), and individual speaker variables such as degree of self-monitoring (Kormos, 1999) and neurobiological factors (Dewaele, 2002). In general, the researchers who have investigated planning have been concerned with its effects on accuracy, syntactic and morphological complexity, and fluency. Generally speaking, it appears that planning time results in greater grammatical complexity (Crookes, 1989; Ortega, 1999; Yuan & Ellis, 2003) and increased fluency (Crookes, 1989; Foster & Skehan, 1996; Mehnert, 1998). The effect of planning on accuracy is more complicated, and depending on the measures taken and the design of the study, mixed results have been obtained. In the current study, we are concerned with speakers' fluency in a variety of tasks, and we also wish to focus on listeners' perceptions of fluency.

Fluency Ratings and Task Differences

Kormos (1999) and Dewaele (2002) have argued that there are long-term settings for fluency that are dependent on individual propensity for a particular speaking style. Despite these settings, however, the research on task type suggests that there are variations in fluency that can be directly attributed to the properties of a given task. Foster and Skehan (1996) examined three tasks commonly found in L2 textbooks: a personal information exchange, a picture narrative, and a collaborative decision-making task. They found that measurements of fluency-related variables on these tasks differed, depending both on the nature of

the task itself and on the availability of planning time. Eijzenberg (2000) has suggested that L2 learners will be perceived as being more fluent in an interaction with a native speaker (NS) than in other situations because they are able to scaffold on the productions of the interlocutor. She argued that in monologic tasks, the cognitive demands on the speaker are greater, and fluency will thus be negatively affected.

Temporal Measures

In this study we examined the relationship between temporal measures and untrained raters' assessments of fluency in low-proficiency L2 speech. Previous studies have involved intermediate to advanced speakers of an L2, whose productions have been objectively assessed for fluency using measures such as speech rate, mean length of run (MLR), and number and duration of pauses. In the current research we were interested in determining whether untrained native listeners could successfully rate the fluency of L2 speakers' speech to the extent that they could differentiate across task types (if such differences appeared). The small number of existing judgment studies have generally involved trained assessors (e.g., Eijzenberg, 1992; Lennon, 1990; Riggenbach, 1991; Wennerstrom, 2000).

An examination of reliability in listeners' judgments of fluency is important in establishing the construct validity of perceived fluency. Many studies of L2 speaker fluency report trained interjudge reliability rates. For example, one study indicates that the ratings of phoneticians and other highly trained listeners were reliable when they judged the fluency of L2 Dutch speech on a 10-point scale (Cucchiarini, Strik, & Boves, 2002). The degree of reliability among less sophisticated listeners, however, is much less clear. Cucchiarini et al. assert that reliability in their own work, which yielded a reliability coefficient of .96 (Cronbach's alpha) for phoneticians, was much higher than in Freed's (1995) and Riggenbach's (1991) studies of ratings made by teachers of English as a second language (ESL).

However, it is unclear that a valid reliability comparison can actually be made, because the latter two studies do not specifically report values for Cronbach's alpha. In fact, if only mean correlations are reported, it may appear that reliability is much lower than in studies that use Cronbach's alpha. Also, because even Cucchiarini et al.'s least sophisticated listeners had received special training on the evaluation of L2 speech, it is not possible to conclude from Cucchiarini et al.'s study how untrained listeners might perform.

In this study we wanted to determine the reliability of ratings from phonetically unsophisticated listeners who received minimal instruction on what to attend to in order to discover whether a common understanding of the notion of fluency would emerge. In earlier work we found that untrained raters were relatively reliable in assessing such features of nonnative speaker (NNS) speech as comprehensibility and accentedness (Derwing & Munro, 1997; Munro & Derwing, 1999). Because a full understanding of listeners' perceptions of fluency requires an examination of judgments from listeners of all types, it is important to examine untrained listeners' sensitivity to temporal factors as well. A key question is whether, given all the other variables in L2 speech, it is possible to focus on fluency while ignoring grammatical errors, unexpected lexical choices, and the like, to the extent that individuals can detect existing fluency differences in the same speaker across tasks. And if raters can differentiate fluency across task types, are their ratings correlated with measurable aspects of fluency typically employed in L2 research (e.g., Foster & Skehan, 1996; Lennon, 1990; Towell et al., 1996), such as speaking rate, MLR, and total number of filled and unfilled pauses? That is, do untrained listeners attend to the temporal variables that are traditionally measured in fluency research? (Note that in a task in which students were required to translate a passage from Dutch to French online, judges' ratings of fluency correlated best with temporal measures [Dewaele, 1994].) These are interesting questions from both methodological and information-processing

perspectives. Given that many other factors influence fluency beyond simple temporal measures (Lennon, 2000), are listeners nonetheless sensitive to simple temporal factors? And if they are, what are the implications for tests that rely on judgments of L2 fluency?

Prosody

In an investigation of the role of pitch and phrasal segmentation, Wennerstrom (2000) showed that prosody affects listeners' perceptions of L2 fluency. She found that L2 speakers who utilized a broad pitch range and who paused in syntactically appropriate places, such as at clause boundaries, were rated by trained judges as more fluent than speakers whose pitch range was limited and who tended to pause in the middle of phrases. In the current study we wanted to determine whether prosody, which clearly contributes to the overall perception of fluency, varies from one task type to another.

Fluency and Voluntary Exposure to English

The relationship between exposure to the L2 and fluency is complex. Although fluency appears to develop with increased exposure to L2 input (Lennon, 1990; Towell et al., 1996), a speaker's initial proficiency may influence the degree of improvement during a stay in an L2 environment. For example, Lapkin, Hart, and Swain (1995) found there were greater gains among students in L2 immersion settings who were initially less proficient than among students with higher proficiency at the outset. Although students in foreign language situations can benefit from extended stays in a country where the L2 they are studying is spoken, few studies have examined the degree or type of exposure in immigrant populations. Will fluency vary depending on whether individuals seek out input, either passively, by watching television or listening to the radio, or actively, by interacting with others in the L2?

Comprehensibility, Accentedness, and Fluency

Finally, we wanted to examine the relationships among perceptions of fluency, comprehensibility, and accentedness. Derwing and Munro (1997) and Munro and Derwing (1999) carried out a series of experiments in which they established that comprehensibility and accentedness are related but partially independent features of nonnative speech. Raters consistently judged accent more harshly than comprehensibility. Perhaps the most interesting finding was that although all “difficult to understand” speech samples were rated as being heavily accented, many heavily accented samples were considered by untrained judges to be relatively easy to understand. In this study, we explored how fluency relates to these two features of L2 speech.

In summary, the research questions we addressed were as follows:

1. Are there differences in fluency ratings across task types?
2. Do fluency ratings of untrained judges correlate with objective temporal measures of fluency?
3. Are goodness-of-prosody ratings related to task type?
4. Is there a relationship between fluency ratings and L2 speakers’ reported exposure to spoken English?
5. What are the relationships among fluency and both comprehensibility and accentedness?

Method

Speakers

The speakers were 20 high-beginner Mandarin-speaking ESL students (7 men and 13 women), ranging in age from 26 to 38 years ($M = 33.4$ years), who were enrolled in full-time LINC classes (introductory ESL courses supported by the federal

government). At the outset, all were assessed as being between Canadian Language Benchmark levels 1 and 3 for speaking and listening (beginner levels). All were professionals in their own country and had entered Canada as skilled workers. None had been in Canada for longer than 6 months prior to the data collection. The speakers completed a language background questionnaire in which they responded to scalar items to measure their voluntary exposure to English. All speakers passed a pure tone hearing screen (250–6,000 Hz at 25 dB).

Listeners

The listeners were 28 English NSs (22 women and 6 men). All were Faculty of Education students enrolled in an undergraduate introductory course in teaching English as a second language at the University of Alberta. They ranged in age from 21 to 52 years ($M = 28.6$). They all reported having normal hearing. None had studied Mandarin, and none indicated ongoing exposure to Mandarin-accented speech.

Speaking Tasks

Recordings were made using high-quality minidisc recorders in a quiet room. The Mandarin speakers completed three speaking tasks during an individual recording session with one of two researchers. The first (Task 1) was an eight-frame picture narrative in which a man and a woman carrying identical green suitcases bumped into each other on a city street. After the incident, they mistakenly exchanged suitcases and later discovered the error after returning to their respective hotel rooms. We gave the ESL speakers 30–45 s to familiarize themselves with the pictures and to ask questions for clarification. (They had seen the same set of pictures once before, 2 months earlier, and had narrated the story both in Mandarin and English at that time.) Once they were ready, the recording device was turned on, and they recounted the story to the researcher. This task was

followed by a 2-min monologue (Task 2) in which the participants talked about the happiest moment in their lives. They were given about 30 s before being recorded to decide which particular moment to talk about. Finally, a conversation with one of the researchers (Task 3) ensued. In this last task, the NNS was directed to ask questions of the researcher about the researcher's happiest moment. Because of the unnaturalness of a researcher inviting questions about his or her personal life, there was some awkwardness in the initial turn; however, a relaxed conversation very quickly ensued.

Stimulus Preparation

We converted all speech samples to high-quality (16-bit) computer audio files so that measurements could be made and so that the stimuli could be conveniently presented to the listeners for judgments. Thirty-second samples were taken from the beginning of the picture narrative and the monologue as soon as the speakers launched into the task (i.e., no initial pauses or false starts were included). We chose to take selections from the beginning to hold the subject matter constant in each narrative. It was our general impression from examining the full transcripts that fluency did not vary to any noticeable degree within tasks. We took 90-s samples from the beginning of the conversations. The longer task length was intended to give listeners the opportunity to adjust to the exchange between the two speakers. We randomized and rerecorded onto a CD the 60 NNS stimuli (one from each of 20 speakers from each of the three tasks) and three samples from two native English speakers for presentation to the listeners.

Listening Task

The listening task was carried out in several small group sessions (for the convenience of the participants) in a quiet room. Before providing judgments, the listeners were told to

pay attention to temporal variables such as filled and unfilled pauses, false starts, and self-repetitions. We told the listeners that we were interested in their perceptions of fluency in terms of the flow and smoothness of speech rather than in terms of overall proficiency. To eliminate familiarity bias, we also showed them the pictures used in the narrative task and told them the topic of the monologue and conversations. (Had we not provided them with this information, they might have judged the first samples differently than later items, simply because they did not know what to expect at the beginning of the task.) The listeners were asked to focus on the person asking the questions in the conversation task (i.e., 20 nonnative English speakers and 1 NS introduced for control purposes). After three practice stimuli, the listeners heard the fully randomized set of 63 speech samples. A researcher paused the stimulus CD for approximately 4–5 s after every sample to allow the listeners to respond. The listeners judged the fluency of each speech sample on a numbered response sheet using a 9-point scale from 1 (*extremely fluent*) to 9 (*extremely dysfluent*). We did not use the type of elaborate descriptors suggested by Fulcher (1996), which are suitable for trained raters in an interview task, because we felt they would be overwhelming for naïve listeners. The listeners in our study were also asked to rate the speech samples for comprehensibility, on a scale from 1 (*extremely easy to understand*) to 9 (*impossible to understand*), and accentedness, on a scale from 1 (*no accent*) to 9 (*very strong accent*). Each listening session lasted about an hour; the participants took a short break halfway through. A few minutes were also used at the end for debriefing.

Goodness-of-Prosody Ratings

We used the same stimulus CD and procedure for the goodness-of-prosody ratings. Three trained judges with extensive experience listening to and evaluating foreign-accented speech independently assessed each speech sample for goodness of prosody, using a 9-point scale from 1 (*native-like prosody*) to

9 (*extremely nonnative prosody*). The raters agreed in advance that they would be listening primarily to intonation and rhythm. After rating the entire set without observing one another's responses, the judges compared their scores and identified 10 cases in which the scores diverged by more than one point. The raters then listened to those samples a second time and negotiated more convergent scores. The ratings were then averaged across the judges for a mean goodness-of-prosody score for each of the stimuli.

Temporal Measures

The speech samples were transcribed in standard orthography by one of the researchers and verified by a second. We made standardized measures using both the transcripts and a waveform editing program. We measured the duration of MLR, filled and unfilled pauses, and speech rate to the nearest millisecond from computer audio files using SoundEdit 16 software. Following Riggensbach (1991), we did not count pauses shorter than 400 ms. We defined MLR as the number of syllables between unfilled pauses of 400 ms or longer, or the number of syllables between nonlexical filled pauses (e.g., "um"). We also counted all syllables, including self-corrections, self-repetitions, false starts, nonlexical filled pauses, and asides. We subtracted all the latter measures from the total number of syllables to compute pruned syllables, which we then divided by the total number of seconds. This measure, pruned syllables per second, provides a useful index of fluency (cf. Mehnert, 1998; Ortega, 1999).

Results

Fluency Ratings and Task Differences

Once we determined that the listeners had been in step while rating the samples (that is, that they had all assigned

ratings of 1 to the NS items), we removed the NS speech samples from the analysis. (Including the NS ratings might have artificially inflated our reliability estimates as well as the intercorrelations among judgments reported below.) To allow comparisons with previous work, we computed interrater reliability in two ways. The mean intercorrelation (r) computed using the method recommended by Hatch and Lazaraton (1991) was .75, indicating an acceptable level of agreement that compares favorably with those obtained in previous studies (e.g., Derwing & Munro, 1997; Munro & Derwing, 2001) and suggests that fluency can be evaluated about as reliably as other dimensions of L2 speech, such as accentedness and comprehensibility. Values of Cronbach's alpha were computed at .95 both for the narrative task and for the monologue task, again indicating an acceptable level of interjudge agreement comparable to that obtained by Cucchiarini et al. (2002) for phonetically trained raters.

On the basis of these findings, we next pooled the listeners' ratings to compute mean ratings for each stimulus item, and we carried out a one-way repeated-measures analysis of variance (ANOVA) on the mean rating data with task (three levels) as the factor. The analysis yielded a significant difference across tasks, $F(2, 54) = 25.18, p < .0001$. Follow-up Bonferroni adjusted t -tests (pairwise comparisons), with the criterion for significance set to $p < .02$, indicated that the NNSs' fluency was judged to be significantly poorer on the picture description task than on either the monologue, $t(27) = 8.30, p < .001$, or the conversation task, $t(27) = 4.20, p < .001$. The monologue and the conversation tasks were not judged to be significantly different from one another in respect to the NNSs' fluency, $t(27) = 1.90, ns$. Cohen's d values for the two significant differences were .75 and .58, respectively, indicating medium effect sizes in both instances.

Measurement Data

Table 1 shows the mean values for five fluency measures of the NNSs' speech on all three tasks. However, we confined our

Table 1

Standardized measures and fluency ratings for the three tasks

Task	MLR	Self-repetition/s	Pause/s	Pruned syllables/s	Speech rate	Mean fluency rating
Picture	4.274	.091	.459	1.379	1.673	5.8
Monologue	4.770	.117	.397	1.662	1.921	5.4
Conversation	4.804	.091	.400	1.607	1.928	5.5

analyses to Tasks 1 and 2 both because the samples were of equivalent length and because we wanted to avoid any influence that the NS interlocutor might have had on the judgments in the third task. The measures that showed the greatest differences between tasks—pausing and pruned syllables—were entered as predictor variables into multiple-regression analyses in which the fluency ratings were the dependent variable. In the picture description task, pausing and pruned syllables together accounted for 69% of the variance in the mean fluency ratings (adjusted $R^2 = .688$, $p < .0001$). For the monologue task, the adjusted R^2 for the same two predictor variables was .648, $p < .0001$, indicating that they explained about 65% of the variance. In neither case did pauses per second make a significant contribution to overall variance accounted for. Although the standardized pause variable was significantly correlated with the fluency ratings in Tasks 1 and 2 ($r = .731$, $p < .01$; $r = .536$, $p < .02$, respectively), its independent contribution to the fluency judgments was not large enough for it to affect the adjusted R -squared value in the analysis involving both variables. We also computed intercorrelations of all the temporal measures (see Table 2) and found relatively high correlations among measures, with the exception of those involving standardized self-repetitions.

Prosodic Goodness

The means for prosodic goodness pooled across listeners for the three tasks were as follows: picture description, $M = 6.61$;

Table 2

Intercorrelations of standardized temporal measures for tasks 1 and 2

	Pause/s	Speech rate	Self-repetition/s	Pruned syllables/s	MLR
Pause/s	1.000	-.701***	-.589**	-.680***	-.689***
		-.750***	-.141	-.752***	-.720***
Speech rate		1.000	.510*	.943***	.777***
			.340	.977***	.876***
Self-repetition/s			1.000	.364	.400
				.184	.246
Pruned syllables/s				1.000	.636**
					.897***
MLR					1.000

Note. Values for task 1 are listed first; task 2 values are below in every case.

* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$.

monologue, $M = 6.57$; and conversation, $M = 6.78$. A repeated-measures ANOVA on the ratings of prosodic goodness yielded no significant differences across tasks, $F(2, 38) = .159$, $p = .8539$, *ns*.

Fluency and Voluntary Exposure to English

In addition, we examined the relationship between fluency ratings and the speakers' reported voluntary exposure to English. We had asked the ESL students to estimate how often they typically had 10-min-or-longer conversations in English over the course of a week outside of school. We also asked them how long they listened to English-language talk radio and television each day (see Table 3 for scales). There was very little variation in the scores for the receptive input (the range on a 5-point scale was 1 to 3; $M = 1.6$); however, we saw sufficient variation across speakers' interactions with both NSs and NNSs in English to support computing correlations (range: 1 to 5 on a 5-point scale; $M = 2.55$). We found a weak but significant

Table 3

Scales for self-reported exposure to L2 input

Type of L2 Input	Exposure				
Passive:					
How much time do you spend watching television/videos in English and/or listening to English language radio each day?	Less than 1 h	2 h	3 h	More than 3 h	
Active:					
How often do you talk in English for 10 min or more?	Never	1-3 times/week	4-6 times/week	Once/day	Several times/day

relationship between fluency judgments and degree of voluntary exposure outside of ESL class on the conversation task, $r(18) = -.483$, $p < .05$, but not on the monologue, $r(18) = -.336$, *ns*, or the picture narrative, $r(18) = -.405$, *ns*.

Comprehensibility, Accentedness, and Fluency

We examined the comprehensibility and accent ratings for all three tasks; interrater reliability scores were $r = .73$ and $r = .72$, respectively. We computed mean comprehensibility, fluency, and accentedness scores for each speaker on each task by pooling the scores over the listeners. We found that comprehensibility was significantly correlated with fluency judgments for Task 1, $r(18) = .825$, $p < .01$; Task 2, $r(18) = .636$, $p < .01$; and Task 3, $r(18) = .873$, $p < .01$. Correlations between fluency and accentedness were somewhat lower, $r(18) = .487$, $p < .05$; $r(18) = .423$, *ns*; and $r(18) = .619$, $p < .01$, respectively.

Discussion

Fluency Ratings and Task Differences

This study advances our understanding of fluency in L2 speech through an examination of untrained listeners' judgments of the productions of low-proficiency Mandarin ESL users. Many previous studies have relied primarily on evaluations from trained listeners, and there are almost no available data on fluency in low-proficiency students. Our listening tasks yielded reliable ratings that reveal patterns in these low-proficiency users that had not previously been investigated. (Clearly, because the results were obtained from only one language group, they are not necessarily generalizable to other language groups.)

The first research question we sought to address was whether there are differences in fluency ratings across task types, as Ejzenberg (2000) suggested. Our results confirmed that the perception of L2 speakers' fluency varied across tasks, as ratings on the picture description task were significantly lower than ratings on either the monologue or the conversation, which did not differ from one another. This finding is similar to that of Foster and Skehan (1996), who found, in comparing three oral tasks in an unplanned condition, that a picture narrative contained more silence than either a personal information exchange task or a collaborative decision-making task. However, Foster and Skehan's participants were less fluent on the collaborative decision-making task (akin to the conversation task reported in this study) than the personal exchange task (essentially monologic) information. They attribute this difference in fluency to cognitive load. The tasks are not exactly comparable to those of the present study or to those in Ejzenberg's (2000) study, in that the conversational interlocutor in Foster and Skehan's task was another ESL student, whereas in the other studies, it was an NS.

The differences among the tasks found in the current study may reflect task-dependent variability in the degree of freedom

the speaker had in choosing lexical items, structures, and content in general, particularly in the absence of planning time. The picture narrative imposed constraints that could not be avoided completely. For example, the L2 students had to describe the collision between the two people carrying suitcases, which many of them found quite difficult, whereas the monologue and conversation tasks may have allowed them to rely on scaffolding and formulaic sequences and to have greater control of the content, through both the telling of familiar stories and the avoidance of possible trouble spots that could lead to communication breakdown. Interestingly, in debriefing, several listeners indicated that they would have rated the samples in the picture description task more negatively if they had not seen the pictures prior to listening to the samples; in that case, the difference among task ratings might have been even greater.

The fact that there was no significant difference in fluency judgments between the monologue and the conversation is not surprising, because both tasks employed the same relatively easy topic, one that the speakers were very likely to have recounted in the past, either in English or their first language (L1). (For example, many people cited the birth of their first child or their marriage as the highlight of their lives.) In fact, the monologue could be viewed as an extended turn in a conversation. Although Foster and Skehan's (1996) personal information exchange task appears to have been quite similar to the monologue in the current study, in that the students reported familiar information (giving directions), it may have been more difficult for the participants than the monologue, because it required clear sequencing, something with which NSs often have trouble.

Measurements of Fluency-Related Phenomena

A second objective of this study was to determine whether fluency ratings are related to temporal measures of speech. We computed multiple-regression analyses with fluency ratings as the dependent variable and found that for both the picture

description and monologue tasks, the measure of standardized pruned syllables was a successful predictor of fluency judgments, accounting for a large portion (65% or more) of the variance. The nature of pruned syllables, essentially a composite measure in which all types of dysfluency are removed, may have accounted for the strong predictive properties of the measure. It should be noted that the intercorrelations among each of the temporal measures were high (see Table 2) and that the fluency measure that appeared to be least related to the others was self-repetition. Self-repetition, however, depending on how it is used by the speaker, could reflect a way to buy time that actually gives an impression of fluency, or it could be perceived as a marker of dysfluent speech (Guillot, 1999). The finding that fluency ratings can be predicted from measurable characteristics of speech further supports the claim that rating data from even untrained listeners reflect properties inherent in the stimuli and are therefore useful in the evaluation of speech samples.

Lennon (2000) has noted that simple temporal measures are somewhat ambiguous and that fluency entails considerably more than being able to speak with few pauses. He refers to Fillmore's (1979) categorization of four dimensions of fluency, which range from the simplest type, as measured here, to the most complex, in which speakers exemplify creativity and wit. Clearly, temporal measures are not the only indicators of fluency, and the causes of dysfluencies may vary. However, in the case of these low-proficiency speakers, all of whom were challenged by a limited lexicon and very little practice speaking in English, temporal measures (on which the listeners were asked to focus) appeared to account for listener judgments relatively well.

Goodness of Prosody

The results of the comparison of goodness-of-prosody ratings and task type indicated that prosodic factors varied little across tasks. Although prosodic accuracy contributes to the

overall impression of fluency (Derwing & Rossiter, 2003; Wennerstrom, 1998, 2000), it is unlikely to vary much over a period of a few minutes, unlike fluency itself, which is subject to a wide range of influences (including knowledge of the content area, linguistic demands, degree of control, availability of formulaic chunks, and interlocutor variables).

Fluency and Voluntary Exposure

The fourth objective of the study was to examine the relationship between fluency ratings and speakers' estimated exposure to English. There was little justification for conducting statistical analyses on the passive input reported by the ESL learners in this study because they spent very little or, in most cases, no time listening to the radio or watching television. However, there was variation across learners in the amount of time they spent conversing in English outside the classroom. It was not surprising that the only significant correlation between fluency and voluntary exposure was on the conversation task, which most closely resembled the students' experiences with speaking English outside the classroom.

Comprehensibility, Accentedness, and Fluency

Our data provided an indication that listeners' perceptions of comprehensibility have a clearer tie to fluency judgments than do their assessments of accentedness. Derwing and Munro (1997) and Munro and Derwing (1999) have repeatedly found that the degree of accentedness of an utterance is only partially related to its comprehensibility. In other words, although speech that is rated nonaccented or lightly accented will almost always be rated easy to understand, and speech that is judged to be difficult to understand will have strong accentedness ratings, heavily accented speech is very often judged as easy to understand. Lennon (2000) has argued that "a good touchstone of acceptable fluency is the degree to which listener attention is

held" (p. 34). Dysfluent speech, whether in L1 or L2 speakers, is disruptive for listeners and is likely to result in a lack of attention. More fluent productions may give the listener the impression that they are easier to understand than less fluent speech, simply because it is easier to attend to language that is not interspersed with hesitation devices, pauses, and false starts. However, it appears that increased fluency is less likely to lead to a perception of reduced accentedness, possibly because accentedness judgments are based more heavily on linguistic phenomena such as segments and prosodic elements. In fact, our finding of a relatively weak relationship between accentedness and fluency in low-proficiency speakers accords well with Burgess's (2001) observations for advanced speakers. Burgess obtained Pearson r values ranging from .40 to .46 on a variety of tasks, including a picture description task identical to the one used here.

An interesting aspect of this study was the low proficiency level of the L2 students. Most research on L2 fluency has been conducted with intermediate and advanced learners, and it has long been assumed that fluency is relatively homogeneous in beginners (that is, that beginners are not fluent). In an analysis of the productions of 100 beginner learners of English, Ranta and Derwing (2000) found that there were significant individual differences in fluency, despite similar low levels of syntactic and lexical knowledge in the L2. The findings in the current study show quite clearly that, very early on in their language studies, L2 learners also show significant differences in fluency depending on task type.

Implications

One of the primary implications of this study for the language classroom is the need to use a variety of tasks that draw upon different skills to enhance fluency, even in pre-intermediate-level classrooms. Rather than limiting oral work to familiar topics, students would benefit from speaking tasks

designed along a continuum of avoidance and control. Tasks that constrain speakers by obliging them to search for unfamiliar words and structures are as important as tasks in which learners can rely on recognizable content either by repeating a task (Bygate, 1996; Nation, 1989) or by scaffolding their productions in interaction (Ejzenberg, 2000). Skehan and Foster (1999) advocate a balance of tasks to ensure that no one aspect of language (accuracy, complexity, and fluency) is overlooked. This advice should perhaps be extended to low-proficiency classrooms, where fluency is traditionally not a high priority.

The learners in this study may also have benefited from a focus on fluency development within the language classroom. Lessons involving formulaic sequences, phrases to buy planning time, and activities designed to encourage facility with paraphrase, appropriate pause placement, and rapid production (see Guillot, 1999) may have helped many of these learners, even though they were at a beginner level of proficiency.

The connection between voluntary exposure to spoken language and L2 fluency suggests that students should be encouraged to use their L2 outside the classroom through contact assignments or service placements. The facilitation of access to other NSs or high-proficiency L2 speakers would benefit learners; some of those in the current study told us that they virtually never had an opportunity to engage in extended conversations with people in English, despite the fact that they lived in an almost exclusively English-speaking part of Canada.

This study also has implications for assessment. As Ejzenberg (2000) has indicated, many oral testing situations involve only an interview, which, given the difference in performance on tasks in this study, may provide an inflated perception of an individual's abilities. If a test is employed to determine fluency for a position that requires a range of oral tasks, then each of those main task types should be included in the assessment. Whereas the raters in this study were not trained, high-stakes tests of fluency generally employ trained individuals. Fulcher's

(1996) caution to use “a data-driven approach to rating scale development” (p. 224) should be taken into account if new tasks are to be included in fluency assessment. A direct comparison of trained and untrained individuals would be most useful in determining the face validity of fluency tests.

Suggestions for Further Research

This study examined a limited range of standard temporal measures and listener ratings. In the future, it would be useful to examine to what extent listeners are influenced by other factors that contribute to the perception of fluency, such as lexical choice and the use of formulaic sequences. Studies that trace the development of fluency across proficiency levels on a variety of task types would also be beneficial. Individual differences should be examined more closely as well, taking into account variables that could affect fluency development, such as learners' L1, native language production rates, degree of exposure to oral input, opportunities for interaction, and willingness to communicate.

It has often been claimed that fluency is a difficult concept to define, yet taxi drivers, shopkeepers, and telephone receptionists, among others, make judgments daily and with ease regarding L2 fluency. The difficulty in pinning down a definition lies in the fact that fluency encompasses so many aspects of language. In many ways, fluency and accent are likely the primary measures of an individual's L2 ability assessed by ordinary interlocutors on the street, regardless of the L2 speaker's actual proficiency. Recalling Lennon's (2000) statement that an individual's fluency is acceptable as long as it holds the listener's attention, it is of crucial importance that we gain a better understanding of how to enhance fluency without sacrificing form.

References

- Burgess, C. S. (2001). *Speaking rate, fluency, and accentedness in the speech of second language learners*. Unpublished doctoral dissertation, Simon Fraser University, Vancouver, British Columbia, Canada.
- Bygate, M. (1996). Effects of task repetition: Appraising the developing language of learners. In J. & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 136–146). Oxford, England: Heinemann.
- Chambers, F. (1997). What do we mean by fluency? *System*, 25, 535–544.
- Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, 11, 367–383.
- Cucchiari, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America*, 111, 2862–2873.
- Derwing, T. M., & Munro, M. J. (1997). Accent, comprehensibility and intelligibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19, 1–16.
- Derwing, T. M., & Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, 13, 1–18.
- Dewaele, J.-M. (1994). Évaluation du texte interprété: Sur quoi se basent les interlocuteurs natifs [How do native-speaker interlocutors evaluate interpreted text]? *Meta*, 39, 78–86.
- Dewaele, J.-M. (2002). Individual differences in L2 fluency: The effect of neurobiological correlates. In V. Cook (Ed.), *Portraits of the L2 user* (pp. 219–250). Clevedon, England: Multilingual Matters.
- Ejzenberg, R. (1992). *Understanding nonnative oral fluency: The role of task structure and discourse variability*. Ann Arbor, MI: University Microfilms International.
- Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287–313). Ann Arbor: University of Michigan Press.
- Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler, & W. S.-Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 85–101). New York: Academic Press.
- Foster, R., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, 18, 299–323.
- Freed, B. F. (1995). What makes us think that students who study abroad become fluent? In B. F. Freed (Ed.), *Second language acquisition in a*

- study abroad context* (pp. 123–148). Philadelphia and Amsterdam: Benjamins.
- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13, 208–238.
- Guillot, M.-N. (1999). *Fluency and its teaching*. Clevedon, England: Multilingual Matters.
- Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. New York: Newbury House.
- Kormos, J. (1999). The effect of speaker variables on the self-correction behaviour of L2 learners. *System*, 27, 207–221.
- Lapkin, S., Hart, D., & Swain, M. (1995). A Canadian inter-provincial exchange: Evaluating the linguistic impact of a three month stay in Quebec. In B. F. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 67–94). Amsterdam: Benjamins.
- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40, 387–417.
- Lennon, P. (2000). The lexical element on spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor: University of Michigan Press.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, 20, 83–108.
- Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, 49 (Suppl. 1), 285–310.
- Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the comprehensibility and accentedness of L2 speech. *Studies in Second Language Acquisition*, 23, 451–468.
- Nation, P. (1989). Improving speaking fluency. *System*, 17, 377–384.
- Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition*, 21, 109–148.
- Ranta, L., & Derwing, T. M. (2000, March). *Accuracy, fluency and learner style*. Paper presented at the meeting of the American Association of Applied Linguistics, Vancouver, British Columbia, Canada.
- Riggenbach, H. (1991). Toward an understanding of fluency: A micro-analysis of nonnative speaker conversations. *Discourse Processes*, 14, 423–441.
- Riggenbach, H. (Ed.). (2000). *Perspectives on fluency*. Ann Arbor: University of Michigan Press.

- Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition*, 14, 357–385.
- Segalowitz, N. (2004, May). *Real-time cognitive processing efficiency and second language fluency acquisition*. Paper presented at the annual meeting of the American Association of Applied Linguistics, Portland, OR.
- Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49, 93–120.
- Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics*, 17, 84–119.
- Wennerstrom, A. (1998). A study of Chinese speakers of English. *Studies in Second Language Acquisition*, 20, 1–25.
- Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102–127). Ann Arbor: University of Michigan Press.
- Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, 14, 85–106.
- Wood, D. (2001). In search of fluency: What is it and how can we teach it? *Canadian Modern Language Review*, 57, 573–589.
- Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, 24, 1–27.

