Does Speaking Task Affect Second Language Comprehensibility?

DUSTIN CROWTHER
Michigan State University
Department
500 West Lake Lansing Road A17
East Lansing, MI, 48823
E-mail: crowth14@msu.edu

PAVEL TROFIMOVICH
Concordia University
Department of Education
1445 de Maisonneuve Blvd., West
Montreal, Quebec
Canada H3G 1M8
Email: pavel.trofimovich@concordia.ca

TALIA ISAACS
University of Bristol
Graduate School of Education
35 Berkeley Square, Bristol
United Kingdom BS8 1JA
E-mail: Talia.Isaacs@bristol.ac.uk

KAZUYA SAITO
Waseda University
School of Commerce
1-6-1 Nishi Waseda,
Shinjuku, Tokyo
Japan 169-8050
Email: kazuya.saito@waseda.jp

<A>ABSTRACT
The current study investigated task effects on listener perception of second language (L2) comprehensibility (ease of understanding). Sixty university-level adult speakers of English from 4 first language (L1) backgrounds (Chinese, Romance, Hindi, Farsi), with 15 speakers per group, were recorded performing 2 tasks (IELTS long-turn speaking task, TOEFL iBT integrated listening/reading and speaking task). The speakers' audio recordings were evaluated using continuous sliding scales by 10 native English listeners for comprehensibility as well as for 10 linguistic variables drawn from the domains of pronunciation, fluency, lexis, grammar, and discourse. In the IELTS task, comprehensibility was associated solely with pronunciation and fluency categories (specifically, segmentals, word stress, rhythm, and speech rate), with the Farsi group being the only exception. However, in the cognitively more demanding TOEFL iBT integrated task, in addition to pronunciation and fluency variables, comprehensibility was also linked to several categories at the level of grammar, lexicon, and discourse for all groups. In both

tasks, the relative strength of obtained associations also varied as a function of the speakers' L1. Results overall suggest that both task and speakers' L1 play important roles in determining ease of understanding for the listener, with implications for pronunciation teaching in mixed L1 classrooms and for operationalizing the construct of comprehensibility in assessments.
<A>END ABSTRACT

Keywords: comprehensibility; task; phonology; fluency; lexicon; grammar; pronunciation learning and teaching

The importance of international trade and education, combined with the ever-growing interest in global popular culture and social media, has underscored the need for speakers to achieve communicative success in multiple languages. Thus, understanding various subcomponents of second language (L2) speaking ability, specifically those contributing to communicative success, has emerged as a chief goal for both language researchers and teachers. With regard to L2 pronunciation, two competing principles have been put forth (Levis, 2005). The first, nativeness, refers to speakers' ability to approximate speech patterns of the target-language community, often measured through accentedness ratings. The second, understanding, denotes speakers' ability to make themselves understood, usually operationalized as comprehensibility and measured through ratings of listeners' ease or difficulty of understanding L2 speech.

Although L2 teachers and students frequently see the acquisition of nativelike, accent-free pronunciation as the ultimate goal of learning (Derwing, 2003; Tokumoto & Shibata, 2011), many proponents of pronunciation have argued for a greater focus on comprehensible speech, as opposed to minimizing first language (L1) influence through accent reduction in instruction and assessment (Derwing & Munro, 2009; Harding, 2013; Levis, 2005). Following this argument, recent research has shown that comprehensible speech is associated with many linguistic factors, spanning the domains of phonology, fluency, lexis, and grammar (Crowther, Trofimovich, Saito, & Isaacs, in press; Saito, Trofimovich, & Isaacs, under review; Trofimovich & Isaacs, 2012). However, the above studies share one serious limitation, namely, a focus on a single task—a picture narrative often used to elicit speech from L2 speakers (e.g., Trofimovich & Isaacs, 2012) despite evidence that linguistic resources needed for speakers to successfully complete a task depend on several variables, including task formality and complexity (Kormos, 2006; Robinson, 2005; Segalowitz, 2010; Skehan, 2009). Furthermore, task effects are a systematic source of variance in L2 speaker performance that could affect scoring outcomes in rated assessment (Upshur & Turner, 1999). Therefore, the goal of this study was to examine whether and to what extent linguistic correlates of comprehensibility depend on the speaking task used to elicit L2 speech. In particular, this study focused on tasks (IELTS long-turn speaking task, TOEFL iBT integrated listening/reading task) that were more appropriate, compared to picture narratives, for use with university students, as these tasks more closely approximate the demands of language use in higher education.

<A>A CASE FOR COMPREHENSIBILITY

The two competing views of L2 pronunciation, termed the nativeness and the intelligibility principles (Levis, 2005), have been the focus of substantial research (see Munro & Derwing, 2011). The nativeness principle highlights nativelike, unaccented L2 pronunciation as a desirable learning and teaching goal, an idea shared by many L2 speakers and their interlocutors,

although attitudes towards nativeness might be slowly changing, particularly in lingua franca contexts (Jenkins, 2013). For example, in a survey of adult immigrants studying English in Canada, Derwing (2003) found that 95% desired to sound like a native speaker, and 59% of the visible minority learners within the group felt that non-accented pronunciation would garner them more respect from Canadians. A similar survey of university students from Tokumoto & Shibata (2011) revealed that 68% of the Japanese and 59% of the Korean participants studying in their home country thought that their accented English limited their ability to communicate effectively. These concerns are not unfounded. Munro (2003) identified stereotyping, harassment, and even occupational loss as ramifications of speaking with an accent, for example, with landlords informing accented speakers that vacant apartments were unavailable or coworkers mimicking accents as a means of ridicule.

In contrast, the intelligibility principle emphasizes L2 speakers' ability to be understood, which is possible even in the presence of a noticeable accent (Derwing & Munro, 1997, 2009). Key to this principle are the constructs of intelligibility and comprehensibility which, though related, reveal different information. Operationalized through scalar ratings, comprehensibility is a measure of listeners' perceived ease or difficulty of understanding L2 speech, while intelligibility is intended to capture listeners' actual understanding, often through the use of orthographic transcriptions of speech (Munro & Derwing, 1999) or comprehension questions related to its content (Hahn, 2004). Despite the focus of intelligibility on listeners' actual understanding, most real-world applications of this construct, which include high-stakes language tests such as IELTS and TOEFL, involve scalar ratings. In this sense, comprehensibility represents a commonplace, practical metric of listener understanding in both research and real-world assessment contexts (Isaacs & Trofimovich, 2012; Levis, 2005).

While individual learners' desire to sound nativelike cannot be ignored, there are factors beyond learners' control that affect their ability to attain accent-free, nativelike pronunciation. In actuality, adult learners rarely pass for native speakers (Bongaerts, 1999; Moyer, 1999) even if they begin learning at an early age (Flege, Munro, & MacKay, 1995), with accented L2 speech generally considered normal and often unavoidable (MacKay, Flege, & Imai, 2006; Major, 2001). And in the rare instances where adult learners do demonstrate native or near-native pronunciation (Bongaerts, 1999; Moyer, 1999), there are usually contributing circumstances, such as amount of exposure, motivation, and type of training, that are unlikely to apply consistently across learners. Comprehensibility, compared to nativelikeness, then appears to be a more realistic L2 learning goal for consistently ensuring communicative success, especially because L2 speakers can be highly comprehensible and intelligible even if they are accented (Kang, Rubin, & Pickering, 2010; Munro & Derwing, 1999).

A focus on comprehensibility, rather than accent reduction or nativelikeness, also seems sensible from a theoretical perspective, particularly within the Interaction Hypothesis (Long, 1996). This view posits that language learning takes place as a result of interactional modifications in conversation. Whenever interlocutors encounter communication breakdowns due to language issues, they often resort to such discourse moves as clarification requests and confirmation checks to resolve the misunderstanding. L2 development is said to occur precisely during these moments, as L2 speakers' attention is drawn to the various linguistic dimensions that may have caused the breakdown (Mackey & Goo, 2007). Assuming that some linguistic dimensions are more likely than others to cause communication breakdowns (Mackey, Gass, & McDonough, 2000), it is important to understand which dimensions are linked to

comprehensibility, in order to help learners notice and repair their nontarget production and ultimately foster their success in L2 communication.

<A>COMPONENTS OF COMPREHENSIBLE SPEECH

If L2 speakers' goal in most real-world contexts is to participate in meaningful interactions, with the intent of making their message clear to interlocutors, prioritizing comprehensibility over nativelikeness is also logical from a practical perspective. However, even if learners and their teachers embrace comprehensibility as a goal, many of them are likely still unclear as to which linguistic factors in L2 speech contribute to making it comprehensible. Initial research analyzing linguistic influences on various measures of understanding, including comprehensibility, focused primarily on phonology and fluency. These studies revealed a variety of factors linked to making L2 speech intelligible and comprehensible, including word stress (Field, 2005), sentence stress (Hahn, 2004), speech rate (Munro & Derwing, 2001), as well as pitch range and pause or syllable length (Kang at al., 2010). Although limited, some evidence exists that poor grammar and inappropriate lexical choices compromise listener understanding (Fayer & Krasinski, 1987; Munro & Derwing, 1999).

More recently, researchers have focused on a combined contribution of multiple linguistic dimensions to comprehensibility. For example, analyzing the speech of 40 French speakers of English, Trofimovich and Isaacs (2012) targeted 19 coded linguistic measures (divided into phonology, fluency, lexis/grammar, and discourse categories). Comprehensibility was best defined through a combination of phonology (word stress), lexis (lexical richness), and grammar (grammatical accuracy). This finding was replicated in a follow-up study looking at the same speakers, but using 11 rated linguistic measures (Saito et al., under review). Once again, comprehensibility was associated with several variables (segmental errors, word stress, lexical richness and appropriateness, grammatical accuracy and complexity). This result was later extended to a sample of 120 Japanese speakers of English (Saito et al., in press) whose comprehensibility was again linked to pronunciation, fluency, lexis, and grammar. Thus, improving comprehensibility requires a focus on several linguistic domains not restricted to pronunciation and fluency.

However, recent research on comprehensibility shares one limitation, namely, a focus on a single task. For example, two of the studies mentioned used a picture-based narrative task to elicit L2 speech (Saito et al., under review; Trofimovich & Isaacs, 2012) while the third targeted individual picture descriptions (Saito et al., in press). Although picture-based tasks are common in L2 speech research (e.g., Derwing, Munro, & Thomson, 2008), allowing for comparisons of results across studies, they are not reflective of real-world contexts, as opposed to tasks speakers may complete, for example, as part of the IELTS or TOEFL iBT proficiency exams, designed to measure their ability to pursue academic degrees in English (Chalhoub–Deville & Turner, 2000). Therefore, picture-based narrative tasks not only overlook learners' real-world communicative needs, they also likely reveal findings which are specific to the task itself.

<A>TASK EFFECT ON COMPREHENSIBILITY

The idea that the language L2 speakers produce is task-dependent is certainly not new. For example, Labov (1966) demonstrated that the production of /r/ by English speakers at several different New York City department stores varied depending on the location of each store along the social hierarchy, and Dickerson (1975) subsequently showed that this variability not only existed among L2 speakers but could also be predicted by task formality. Over the past decades, numerous theoretical proposals have been proposed to explain how various aspects of task

contribute to speakers' L2 output, including Tarone's Capability Continuum Model (1983), Skehan's Trade-Off Hypothesis (1998, 2009), and Robinson's Cognition Hypothesis (2001, 2005). Tarone, for instance, proposed that the linguistic variables in L2 speech shift depending on task formality, ranging between a pidgin-like vernacular and a careful targetlike style. Skehan argued that cognitively complex tasks create competing processing demands, leading to tradeoffs between linguistic complexity and accuracy, while Robinson suggested that complex tasks, as opposed to cognitively less demanding ones, elicit more elaborate language as speakers strive to meet the greater demands placed upon them. Regardless of the theoretical position taken, different tasks impact multiple aspects of L2 output, from segments and prosody (Tarone) to grammar and lexicon (Robinson, Skehan). And because multiple linguistic dimensions are linked to comprehensibility (Saito et al., in press; Trofimovich & Isaacs, 2012), different tasks will certainly draw on different linguistic resources as L2 speakers try to make themselves understood.

With respect to L2 pronunciation, there is currently limited research targeting speech output in different tasks and no research focusing on comprehensibility. In terms of fluency, for example, Derwing, Rossiter, Munro, and Thomson (2004) showed that L2 speakers receive higher fluency ratings in monologue- or dialogue-based tasks than in a picture-narrative task. L2 speakers are also perceived as being more fluent in dialogue- than in monologue-based tasks (Ejzenberg, 2000), revealing a hierarchy in task types, with dialogue-based tasks producing the most and picture narratives eliciting the least fluent speech in terms of listeners' perceptions. As picture narratives require speakers to describe objects and actions depicted in each image, perceived disfluencies in this task may be due to a search for vocabulary that learners might not know (Hilton, 2008), which should matter less in unstructured monologue- and dialogue-based tasks. In addition, picture narratives would suggest the use of a culturally-bound storytelling discourse structure, which is not a constraint imposed by extemporaneous speech tasks, where speakers are at greater liberty to select the lexical items they wish to use (Martin & Rose, 2003).

Beyond fluency, task impacts both accuracy and complexity of L2 speech. For accuracy, Yuan and Ellis (2003) found that L2 speakers who had sufficient time for task completion, compared to those who did not, could use appropriate vocabulary with correct grammar, likely because planning time with little communicative pressure allowed speakers to monitor their speech. Similarly, L2 pronunciation research has shown that the production of /θ/ by Chinese speakers of English follows the same hierarchy as that established by Dickerson (1975), with read-aloud tasks eliciting more accurate production of L2 segments than more spontaneous tasks, such as storytelling and interviews (Rau, Change, & Tarone, 2009). With respect to complexity, for example, Skehan and Foster (1997) found that narrative-based tasks generated more accuracy and less complexity in L2 utterances, whereas tasks that required some form of decision (i.e., giving advice to people with personal problems) generated more complexity but less accuracy. Robinson (2001) also showed L2 speakers producing greater lexical variation in a more complex version of a map task (giving directions based on a large map of an unfamiliar location), compared to its less complex version (smaller map of a well-known location). In essence, task complexity appears to lead to accuracy-complexity tradeoffs in vocabulary and grammar while the availability of planning time positively impacts speech accuracy.

<A>THE CURRENT STUDY

Clearly, when considering linguistic measures that underlie comprehensibility, the importance of task cannot be ignored, especially because many linguistic dimensions of speech shown to vary by task (individual segments, aspects of fluency, lexicogrammar) have also been

linked to comprehensibility (Trofimovich & Isaacs, 2012). Given that prior research on comprehensibility has been mostly limited to a single task, it remains to be seen whether these associations will hold across tasks. Therefore, the current study addressed this issue by investigating the effect of speaking task on the relationship between comprehensibility and various linguistic dimensions of L2 speech. The following research question guided the study:

RQ. Do relative contributions of various linguistic dimensions of L2 speech to comprehensibility vary as a function of the speaking task performed?

To address this question, the speech of 60 L2 English learners completing two tasks was audio recorded and then rated for comprehensibility and analyzed for 10 linguistic dimensions by 10 native-speaking listeners. Because all speakers were students at an English-medium university, the tasks required them to use skills targeted by the two tests most frequently used for university admission purposes in North America (TOEFL iBT, IELTS), which ensured that the speakers engaged in tasks thought to predict their ability to use academic English and reflect at least some demands of their daily English use arguably to a greater extent than cartoon-style picture narrative tasks. Most importantly, the tasks differed along several dimensions, such as planning time, familiarity of task elements, as well as causal and intentional reasoning, and perspective-taking.

<A>METHOD

<B>Participants

<C>*Speakers*. Participants were 60 L2 speakers drawn from an unpublished corpus of 143 speakers from 19 linguistic backgrounds completing five tasks (Isaacs & Trofimovich, 2011). All speakers were in their first semester of studies as undergraduate (29) or graduate (31) students at an English-medium Canadian university. They were assigned to four groups (n = 15) based on native language (L1) background to disentangle task effects from possible L1 influences on comprehensibility, since linguistic dimensions of L2 speech are known to be L1-specific (e.g., Eckman, 2004; Escudero & Boersma, 2004; Flege, 2003). The groups included speakers of Mandarin Chinese, Hindi/Urdu, Farsi, and Romance languages. The Chinese, Hindi/Urdu (Dravidian languages, which differ in script only), and Farsi groups represented the three largest cohorts in the corpus (with 15, 17, and 32 speakers, respectively). The Romance group included all speakers of French (10) and Spanish (5), which come from the same language family and share a syllable-timed rhythm (Jun, 2005). The four L1s also crucially differ in their segmental inventories (e.g., Duanmu, 2007; Shackle, 2001; Wilson & Wilson, 2001) as well as prosody, particularly in terms of rhythm, thus allowing for direct comparisons between the speakers of syllable-timed French tested by Trofimovich and Isaacs (2012) and the speakers of non-Romance syllable-timed Hindi (Shackle, 2001), stress-timed Farsi (Jun, 2005), and tonal Chinese (Jun, 2005). Apart from gender composition in the Hindi/Urdu group (which reflected the gender distribution of Hindi/Urdu speakers in the larger university community), the speakers were matched as much as possible for several background variables, summarized in Table 1. Although minor differences existed across groups in self-ratings of speaking and listening and self-reported L2 use, there were no significant differences across the groups in TOEFL or IELTS overall scores (i.e., objective measures of proficiency), Fs < 1.67, p > .20, or listening and speaking subscores, Fs < 2.23, p > .11.

<INSERT TABLE 1 HERE>

TABLE 1

L2 Speakers' Background Characteristics by L1 Group (Means and Standard Deviations)

Background Variable	Chinese	Hindi/Urdu	Farsi	Romance
Gender (m/f)	6/9	14/1	9/6	9/6
Age	22.5 (2.9)	23.5 (2.0)	25.2 (2.4)	21.4 (3.3)
Years of English study	10.3 (2.9)	14.3 (6.0)	8.5 (4.8)	11.1 (4.3)
Years in Canada	0.7 (0.3)	0.4 (0.2)	0.4 (0.2)	0.6 (0.3)
TOEFL iBT total score	84.8 (5.9)	92.6 (4.8)	87.8 (7.1)	82.0 (8.5)
IELTS total score	6.3 (0.5)	6.7 (0.6)	6.8 (0.4)	7.0 (0.5)
Speaking ability[a]	5.6 (1.2)	7.1 (1.1)	5.7 (1.0)	6.4 (1.1)
Listening ability[a]	5.9 (1.4)	8.0 (0.8)	6.9 (0.9)	7.6 (1.0)
English use at home[b]	17.0 (16.9)	40.0 (26.5)	21.0 (34.1)	29.3 (32.8)
English use at school[b]	72.7 (21.5)	83.3 (20.6)	50.0 (30.5)	79.3 (27.1)

Note. [a]Self-rating on a 1–9 scale (1 = *extremely poor*, 9 = *extremely fluent*). [b]Self-rating on a 0–100% scale.

<C>*Raters*. Ten native English speakers, with a mean age of 32.7 years (25–56), participated as experienced raters. They were current students (9) or recent graduates (1) of advanced degrees in applied linguistics (7 MA, 3 PhD), and had on average 6.6 years of L2 teaching experience (1–23). Experienced raters were chosen over inexperienced ones because the former have shown greater consistency in evaluating complex and less intuitive linguistic categories in a similar rating task (Saito et al., under review). All raters were raised in English-speaking homes, with at least one parent a native speaker, and reported using English 89% of the time daily (80–100%). Because listeners' familiarity with L2 speech can impact their judgments, only the raters who reported high familiarity with accented English were selected.[1] As a group, they estimated their experience with accented English at 8.6 (7–9) on a 9-point scale (1 = *not at all familiar*, 9 = *very familiar*).

<*B*>*Speaking Tasks*

The speakers in the original corpus completed five tasks, administered in eight randomized orders, distributed equally across speakers: a read-aloud task, a picture narrative task, an IELTS long-turn speaking task, a TOEFL iBT integrated task, and a Test of Spoken English graph-based interpretation task. Of these, two were selected for the comparison of task effects – the IELTS long-turn task (hereafter, the IELTS task) and the TOEFL iBT integrated task (hereafter, the TOEFL task). First, both tasks are part of high-stakes instruments used to make critical decisions about non-native applicants' ability to pursue academic studies in higher education (Chalhoub–Deville & Turner, 2000). Considering the high stakes attached to language performance in both tasks in real-world assessment settings, it is therefore important to investigate how differences between the two tasks may affect L2 comprehensibility. Second, according to Robinson's (2007) framework of task classification, the tasks drew on different sets of cognitive resources, with the TOEFL task being cognitively more complex than the IELTS task. Briefly, the IELTS task required reference to some familiar elements and spatial locations (+ few elements, + known spatial relationships), without the need for reasoning about causal events and relationships or people's intentions and beliefs (− causal and intentional reasoning), and no requirement to depart from a first-person narrative (− perspective-taking). In contrast, the TOEFL task targeted unfamiliar factual and spatial information (− few elements, − known spatial relationships) and required reasoning and perspective-taking (+ causal and intentional reasoning, + perspective-taking).

Third, the two tasks differed in amount of planning (both before and during the task). Whereas the IELTS task allowed speakers to take notes before beginning their response (1 min) and to speak without much time pressure (2 min), the TOEFL task imposed a time limit for both pre-task (30 s) and online (1 min) planning. Finally, the two tasks differed in the extent to which they involved a human interlocutor. The IELTS task exemplified a dialogic, direct test which assessed speaking through face-to-face oral communication with an interviewer, although the interviewer's role was limited. In contrast, the TOEFL task was a monologic, semi-direct test, which included a machine-mediated assessment involving a test-taker speaking into a recording device (Ellis, 2001; Qian, 2009). In sum, the tasks differed in the cognitive load they imposed on the speaker, with the IELTS task being a less demanding face-to-face speaking task, compared with the TOEFL task, which not only placed higher cognitive demands on speakers but also engaged them in monologic performance, without the need to interact with a human interlocutor.

Both tasks used publically-available versions of sample test materials (Educational Testing Service, 2006; IELTS, 2009; Jakeman & McDowell, 2008). The IELTS task assesses a

test-taker's ability to speak from a written prompt. Following IELTS procedures, the speakers received a card with their assigned topic and suggestions of possible discussion points. Two task prompts were used (*describe a sports event you enjoyed watching* or *describe a job you would like to do in the future*), with half of the speakers randomly assigned to one of the two prompt conditions. The speakers had up to 1 min to consider their response and take notes before speaking for 1–2 minutes. The interviewer did not engage in a conversation with a speaker but followed up each response with 1–2 questions, based on IETLS procedures (e.g., *Do you play this sport yourself?* for the prompt about a favourite sports event).

The TOEFL task assesses a test-taker's ability to integrate information from multiple sources (listening and reading) and present it coherently. Following TOEFL iBT procedures, audio and visual prompts were presented through a computer-based interface. The speakers first had 45 seconds to read a passage (93–105 words); they then listened to an audio recording of a lecture (80–90 seconds) related to the passage. They were then asked to draw on examples from both the reading and the audio to respond to a question related to the content from the two sources of input. Two task versions were used, with the topics of audience effects in psychology or explaining behavior in sociology, and half of the speakers were randomly assigned to each. The speakers had 30 seconds to prepare their response and 1 minute to speak, moderated by an audio recorded examiner used for all participants.

During data collection as part of the larger project, speakers' output was recorded using a Plantronics (DSP–300) microphone. The 120 target recordings (60 per task) were prepared for analysis by matching peak amplitude across files, removing all initial fillers and false starts, and then editing down all files to the initial 30 seconds of speech produced, in line with prior research using 20–60 second recordings for listener judgments (e.g., Derwing et al., 2008). All files were orthographically transcribed by a trained research assistant and subsequently verified by a second researcher. The resulting audio files and transcripts served as stimuli for raters' judgments of comprehensibility and their ratings for 10 linguistic categories.

<B>Rating Procedure

All ratings were collected as part of a larger project evaluating speaker performance across multiple tasks. The project took place over four individual 2-hour sessions (with breaks), occurring within a three-week span. Session 1 involved rating accentedness (reported elsewhere) and comprehensibility for all audio recordings. Session 2 and part of Session 3 were dedicated to rating audio recordings for five phonology- and fluency-based categories. The remainder of Session 3 and Session 4 was used to evaluate transcripts for five lexical, grammatical, and discourse categories. In all sessions, audio recordings and transcripts were blocked and counterbalanced by task (e.g., Task 1-2-3; 2-3-1, etc.), with audio recordings or transcripts presented to each rater in a unique randomization.

All ratings were collected through 1000-point scales run in a computer-based MATLAB interface developed by Saito et al. (in press). Each scale included a free-moving slider on a horizontal plane, with the leftmost (negative) end corresponding to "0" and the rightmost (positive) end corresponding to "1000". No numeric labels or interval markings were included, other than brief description of each category's endpoints (see Online Supporting Documentation). At the start of each session, raters received training on the rating interface and each relevant linguistic category for that session. They then proceeded to perform four practice judgments (either through listening to audio files or viewing transcripts), which were discussed with the researcher to ensure an accurate understanding of each measure. The raters were encouraged to use the entire scale range, and were informed that even the slightest shift of the

slider (which was initially set in the middle) might represent a fairly large change in rating. All relevant scales in each session (e.g., five categories for audio rating) were visible simultaneously. Before proceeding to the next recording (or transcript), the raters were allowed to adjust their judgments on all visible scales until they felt satisfied. To assess rater understanding and comfort with each rated category, they used 9-point scales to self-rate their understanding of each category (1 = *I did not understand at all*, 9 = *I understand this concept well*) and comfort in using each (1 = *very difficult*, 9 = *very easy and comfortable*). As a group, they estimated their understanding at 8.3 (7.8–8.7) and their comfort at 7.8 (7.2–8.3), suggesting that they understood all constructs following training and could apply them easily.[2]

<B>*Rated Categories*

<C>*Comprehensibility*. Comprehensibility was defined as the degree of ease or difficulty in raters' understanding of L2 speech (see Online Supporting Documentation). Consistent with previous research on listener-based ratings of comprehensibility, the raters listened to each recording once before making their decision.

<C>*Phonology and Fluency*. The raters evaluated each audio recording for the following five segmental, prosodic, and temporal categories (shown in full in Online Supporting Documentation):

1. Segmental errors (1 = *frequent*, 1000 = *infrequent or absent*), defined as errors in the pronunciation of individual consonants and vowels within a word (e.g., *intelesting* instead of *interesting*; *gud* instead of *good*), as well as any segments erroneously deleted from or inserted into words (e.g.,*'ospital* instead of *hospital*; *sutrength* instead of *strength*).
2. Word stress errors (1 = *frequent*, 1000 = *infrequent or absent*), defined as errors in the placement of primary stress (e.g., *bal-co-NY* instead of *BAL-co-ny*, where capitals designate primary stress) or the absence of discernible stress, such that all syllables receive equal prominence (e.g., *bal-co-ny*).
3. Intonation (1 = *unnatural*, 1000 = *natural*), defined as appropriate pitch moves that occur in native speech, such as rising tones in yes/no questions (e.g., *Did you see the game last night* ↑) or falling tones at the end of statements (e.g., *No, I was too busy* ↓).
4. Rhythm (1 = *unnatural*, 1000 = *natural*), defined as the difference in stress (emphasis) between content and function (grammatical) words. For instance, in the sentence *My SISTER WORKS in an OFFICE*, the words *sister*, *works*, and *office* are content words and therefore are stressed more than the words *my*, *in*, and *an*, which are grammatical words featuring reduced vowels.
5. Speech rate (1 = *too slow or too fast*, 1000 = *optimal*), defined as a speaker's overall pacing and the speed of utterance delivery.

Because judgments of phonology and fluency likely require an in-depth analysis of the speech signal, the raters had the option to listen to the same file multiple times to ensure they were confident in the final rating.

<C>*Lexicon, Grammar, and Discourse*. To remove pronunciation and fluency as possible confounds in judgments of lexis, grammar, and discourse, the raters evaluated written transcripts of the audio files (Crossley, Salsbury, & McNamara, 2014). The transcripts were edited to remove hesitation markers (e.g., um, uh), spelling clues signaling pronunciation-specific errors (e.g., *when*, although pronounced as *ven*, was still spelt as *when*), and punctuation to avoid transcriber influence (Ochs, 1979). The raters evaluated written transcripts for the following five lexical, grammatical, and discourse categories (see Online Supporting Documentation):

6. Lexical appropriateness (1 = *many inappropriate words used*, 1000 = *consistently uses appropriate vocabulary*), defined as the speaker's choice of words to accomplish the task. Poor lexical choices include incorrect, inappropriate, and non-English words (e.g., *She was quite happy when she was old enough to read the papernews*).
7. Lexical richness (1 = *few, simple words used*, 1000 = *varied vocabulary*), defined as the sophistication of the vocabulary used by the speaker. Simple words with little variety correspond to poor lexical richness (e.g., *The young man went for a walk; his cat went for a walk after him*, compared to *The young man went for a walk and was followed by his cat*).
8. Grammatical accuracy (1 = *poor grammar accuracy*, 1000 = *excellent grammar accuracy*), defined as the number of grammar errors made by the speaker. Examples included errors of word order (e.g., *Where we are going?*), morphology (e.g., *He eat a big breakfast every morning*), and agreement (e.g., *I watched two game last week*).
9. Grammatical complexity (1 = *simple grammar*, 1000 = *elaborate grammar*), defined as the sophistication of the speaker's grammar. Grammatical complexity is low if the speaker uses simple, coordinated structures without embedded clauses or subordination (e.g., *The girl asked the man for his seat; she wore high heels*, compared to *The girl wearing high heels asked the man for his seat*).
10. Discourse richness (1 = *simple structure, few details*, 1000 = *detailed and sophisticated*), defined as the richness and sophistication of the utterance content. Discourse richness is low if the entire narrative is simple, unnuanced, bare, and lacks sophisticated ideas or details, but high if the speaker produces several distinct ideas or details so that the statement sounds developed and sophisticated.

As with phonology and fluency judgments, the raters were given as much time as they needed with each transcript to allow for accurate judgments.[3]

<*B*>*Data Analysis*

The 10 raters showed high consistency in their rating. For comprehensibility, reliability values (Cronbach's alpha) were high in both tasks (a_{IETLS} = .91, a_{TOEFL} = .92). For individual linguistic categories, which presumably reflect less intuitive and more complex judgments compared to comprehensibility ratings, reliability values also exceeded the benchmark of .70–.80 in both tasks (Larson–Hall, 2010), as shown in Table 2. The scores were thus considered sufficiently consistent and were averaged across the 10 raters to derive a single mean score for each rated category.

<INSERT TABLE 2 HERE>

TABLE 2
Rater Consistency (Cronbach's alpha) for 10 Rated Linguistic Categories

Linguistic Dimensions	IELTS	TOEFL
Segmental errors	.93	.93
Word stress errors	.86	.84
Intonation	.87	.87
Rhythm	.84	.88
Speech rate	.85	.91

Lexical appropriateness	.84	.84
Lexical richness	.85	.90
Grammatical accuracy	.87	.87
Grammatical complexity	.89	.90
Discourse richness	.90	.90

<A>RESULTS

<B>Comprehensibility

The first analysis targeted task- and group-based differences in comprehensibility ratings (summarized in Table 3). These ratings were submitted to a two-way analysis of variance (ANOVA) with group (Chinese, Hindi/Urdu, Farsi, Romance) as a between-subjects factor and task (IELTS, TOEFL) as a within-subjects factor. The ANOVA yielded a significant main effect of group $F(3, 56) = 9.40$, $p < .0001$, $\eta_p^2 = .34$, and task, $F(1, 56) = 9.90$, $p = .003$, $\eta_p^2 = .15$, but no significant two-way interaction, $F(3, 56) = 1.08$, $p = .36$, $\eta_p^2 = .06$. Tests of simple main effects (Bonferroni-corrected $a = .007$) further showed that the speakers were overall more comprehensible in the IELTS task than the TOEFL task ($p = .003$), although the effect size was small ($d = .35$), and that the Chinese group was overall less comprehensible, with large effect sizes ($d = 1.01$–1.73), compared to the remaining three groups ($p < .001$) which did not differ in comprehensibility.

<INSERT TABLE 3 HERE >

TABLE 3
Means (Standard Deviations) for Comprehensibility Ratings (1000-Point Scale)

L1 Group	IELTS Long-Turn	TOEFL iBT Integrated	Total
Chinese	556 (127)	467 (101)	511 (91)
Hindi/Urdu	701 (150)	662 (145)	681 (134)
Farsi	666 (88)	656 (117)	661 (81)
Romance	754 (141)	689 (166)	722 (145)
Total	669 (145)	618 (159)	644 (139)

<B>Ratings of Linguistic Categories

The next analysis focused on the relationship between comprehensibility and the 10 rated linguistic variables. First, the linguistic scores for all speakers were submitted to an exploratory Principal Component Analysis (PCA) with Oblimin rotation (carried out separately for each task) to determine whether the 10 linguistic variables showed any underlying patterns.[4] Despite a low sample size ($N = 60$), the Kaiser–Meyer–Oklin value exceeded the required .60 for sampling size for both tasks (IELTS = .86, TOEFL = .91), indicating excellent factorability of the correlation matrix (Hutcheson & Sofroniou, 1999). Similarly, Bartlett's test of sphericity for the IELTS task, $\chi^2(45) = 700.57$, $p < .0001$, and the TOEFL task, $\chi^2(45) = 822.95$, $p < .0001$, indicated that

correlations between variables were sufficient for PCA.[5] As shown in Table 4, the analyses revealed the same two underlying factors for both tasks. Factor 1, labelled "Pronunciation," encompassed the five pronunciation and fluency variables. Factor 2, labelled "Lexicogrammar," included the five vocabulary, grammar, and discourse variables. For both tasks, then, the 10 linguistic variables patterned along two separate dimensions (pronunciation, lexicogrammar), accounting in total for a similar amount variance in both tasks (IELTS = 82%, TOEFL = 83%).

<INSERT TABLE 4 HERE>

TABLE 4
Summary of a Two-Factor PCA Solution for 10 Rated Linguistic Variables by Task

PCA Factors	IELTS	TOEFL
Factor 1 (Pronunciation)	Intonation (.95), segmental errors (.93), rhythm (.93), word stress errors (.89), speech rate (.61)	Segmental errors (1.01), intonation (.97), word stress errors (.90), rhythm (.88), speech rate (.64)
Factor 2 (Lexicogrammar)	Lexical richness (.95), discourse richness (.95), grammatical complexity (.91), lexical appropriateness (.87), grammatical accuracy (.82)	Lexical richness (1.02), discourse richness (.99), grammatical complexity (.98), grammatical accuracy (.85), lexical appropriateness (.73)

Note. All eigenvalues > 1.

The pronunciation and lexicogrammar PCA scores, derived using the Anderson–Rubin method of obtaining non-correlated factor scores (Field, 2009), were then submitted as predictors to two separate stepwise multiple regression analyses to examine their contribution to IELTS and TOEFL comprehensibility scores. As shown in Table 5, the two factors together accounted for a high proportion of shared variance in each tasks (IELTS = 74%, TOEFL = 88%). However, both lexicogrammar (IELTS = 14%, TOEFL = 17%) and pronunciation (IELTS = 60%, TOEFL = 71%) seemed to weigh more heavily in the TOEFL than in the IELTS task.

<INSERT TABLE 5 HERE>

TABLE 5
Summary of Multiple Regression Analyses Using Pronunciation and Lexicogrammar Factors as Predictors of Comprehensibility by Task

Task	Predictor Variables	*Adjusted* R^2	R^2 *change*	$F(1, 59)$	*p*
IELTS	Pronunciation	.60	.60	88.89	.0001
	Lexicogrammar	.74	.14	82.61	.0001
TOEFL	Pronunciation	.71	.71	143.82	.0001
	Lexicogrammar	.88	.17	203.75	.0001

Note. The variables entered into the regression equation were the two factors obtained in the PCA analyses summarized in Table 4.

<B>Task Effect and Linguistic Correlates of Comprehensibility

In the next analysis, two sets of partial correlations targeted the relative associations of the individual pronunciation and lexicogrammar variables with comprehensibility, separately for each task and group. For correlations between comprehensibility and the five pronunciation variables, the five lexicogrammar variables were partialled out. In turn, when examining the lexicogrammar-comprehensibility links, the five pronunciation variables were partialled out. Comprehensibility scores in the IELTS task (shown in Table 6) were primarily linked to pronunciation variables, specifically segmental errors and rhythm (with Farsi speakers being the exception). In the TOEFL task (shown in Table 7), comprehensibility scores for all L1 groups were associated with a broader range of variables, spanning both the pronunciation (primarily segmental errors) and lexicogrammar (especially lexical appropriateness and grammatical accuracy and complexity) factors.

<INSERT TABLE 6 HERE>

TABLE 6
Partial Correlations Between Comprehensibility and 10 Linguistic Categories in the IELTS Long-Turn Task

Category	Chinese	Hindi/Urdu	Farsi	Romance
Segmentals[a]	.78*	.75*	.58	.80*
Word stress[a]	.61	.62*	.62	.57
Intonation[a]	.36	.57	.54	.61
Rhythm[a]	.66*	.69*	.81*	.46
Speech rate[a]	.70*	.54	.55	.61
Lexical appropriateness[b]	.54	.46	.88*	.62
Lexical richness[b]	.52	.39	.81*	.43
Grammatical accuracy[b]	.27	.48	.78*	.46
Grammatical complexity[b]	.49	.42	.48	.52
Discourse richness[b]	.54	.60	.78*	.43

Note. *$p < .05$. [a]Partialled-out variables include lexical appropriateness and richness, grammatical accuracy and complexity, and discourse richness. [b]Partialled-out variables include vowel and consonant errors, word stress, intonation, rhythm, and speech rate.

The final analysis sought to confirm the relationships identified through partial correlations by investigating whether each group's performance in the IELTS and TOEFL tasks could be explained through speaker scores for the 10 linguistic variables. Two discriminant analyses were conducted, separately for the IELTS and TOEFL tasks, with the goal of predicting which L1 group each of the 60 speakers belonged to, based on the ratings they had received for each of the 10 linguistic variables.[6] In the IELTS task, there were three significant functions which individually differentiated the four L1 groups, $\Lambda = .19$, $\chi^2(30) = 86.22$, $p < .0001$. The first

function involved lexical richness (r = .44) and discourse richness (r = .47). The second function was associated with rhythm (r = .59) and speech rate (r = .45). The third function was linked to segmental (r = .60) and word stress (r = .46) accuracy. In contrast, for the TOEFL task, there was a single function distinguishing between the four groups, $\Lambda = .32$, $\chi^2(30) = 58.75$, $p < .0001$. This function included all 10 linguistic categories (r =.49–.78). In essence, the 60 speakers could be discriminated as belonging to their respective L1 groups through their *individual* ratings of (a) lexical and discourse richness, (b) rhythm and speech rate, or (c) segmental and word stress accuracy in the IELTS task. However, all 10 linguistic variables *in combination* contributed to discriminating across the 60 speakers in the TOEFL task.

<INSERT TABLE 7 HERE>

TABLE 7
Partial Correlations Between Comprehensibility and 10 Rated Linguistic Categories in the TOEFL iBT Integrated Task

Category	Chinese	Hindi/Urdu	Farsi	Romance
Segmentals[a]	.76*	.40	.88*	.82*
Word stress[a]	.48	.46	.61	.69*
Intonation[a]	.22	.70*	.87*	.57
Rhythm[a]	.39	.24	.94*	.67*
Speech rate[a]	.36	.60	.68*	.88*
Lexical appropriateness[b]	.63*	.87*	.69*	.66*
Lexical richness[b]	.62*	.59	.66*	.32
Grammatical accuracy[b]	.79*	.69*	.74*	.38
Grammatical complexity[b]	.76*	.71*	.73*	.30
Discourse richness[b]	.40	.76*	.85*	.22

Note. *$p < .05$. [a]Partialled-out variables include lexical appropriateness and richness, grammatical accuracy and complexity, and discourse richness. [b]Partialled-out variables include vowel and consonant errors, word stress, intonation, rhythm, and speech rate.

<A>DISCUSSION

The research question of the current study asked whether the relationship between listener ratings of comprehensibility and various linguistic dimensions in L2 speech depend on speaking tasks varying in cognitive demands. Speakers from four different L1 backgrounds completed two tasks (IELTS, TOEFL) that differed in their cognitive requirements. Compared to the IELTS task, the TOEFL task required speakers to consider a greater number of elements and employ more reasoning, and placed greater time constraints on speakers in formulating their response. Results overall suggested that task serves an important role in determining which linguistic variables are linked to comprehensibility. In the IELTS task, comprehensibility was associated solely with pronunciation and fluency categories (specifically, segmentals, word stress, rhythm, and speech rate) for three of the four groups, with only the Farsi group demonstrating associations with lexicon, grammar, and discourse (see Table 6). However, in the

cognitively more demanding TOEFL task, in addition to pronunciation and fluency variables, comprehensibility was also linked to several categories at the level of grammar, lexicon, and discourse for all groups (see Table 7). Although, predictably, the four groups featured slightly different patterns of associations, likely determined by cross-linguistic differences between the speakers' L1 and English (as discussed subsequently), the pattern was clear: In a cognitively more demanding task, compared to a simpler one, ease of understanding was based on appropriate and rich vocabulary, accurate and complex grammar, and rich discourse structure, in addition to nativelike pronunciation and fluency.

<B>Task Complexity and Comprehensibility

For all speakers as a group, comprehensibility was rated higher in the IELTS task than in the TOEFL task. The IELTS task did not require speakers to elaborate on causal relationships, incorporate multiple perspectives, or interpret any complex meanings, as speakers were asked to express an opinion about a straightforward topic, which they likely had numerous prior opportunities to consider and discuss (i.e., future job or favourite sport). In addition, there were few steps involved in the task, such that speakers simply read a prompt, spent a minute planning, and then responded to it. Robinson's Cognition Hypothesis (2001, 2005) predicts that cognitively more challenging tasks should lead to more elaborate language so that speakers can meet increased task demands. Thus, if task demands are not high and task content is highly familiar, speakers face little need to rely on rich, complex vocabulary and grammar forms in generating their response. Instead, they may choose to draw upon familiar words and structures, possibly relying on previous experience discussing similar topics when formulating their utterances. It is likely that this linguistic and thematic freedom, where lexical and grammatical choices may be considered safe, contributes to ease of understanding for the listener being mostly linked to pronunciation and fluency aspects of L2 speech.

Compared to the IELTS long-turn task, the TOEFL integrated task requires speakers not only to rely on receptive language skills (reading, listening or both) but also to interpret multiple sources of information and subsequently integrate this information into a coherent response. To accomplish this task successfully and effectively synthesize the various sources of information, speakers must be able to produce elaborate language, which creates more opportunities for lexical and grammatical errors to occur and for discourse structure to suffer, thus leading to a greater impact of these linguistic variables on comprehensibility. This could be seen in the greater amount of variance explained by the lexicogrammar factor in the regression analysis in the TOEFL task (17%) than in the IELTS task (14%), along with the greater number of significant correlations between comprehensibility and lexicogrammar variables in the TOEFL task, compared to the IELTS task (cf. Tables 6 and 7). In fact, in the TOEFL task, lexical appropriateness was linked to comprehensibility for all four groups, grammatical accuracy and complexity for three groups, and lexical and discourse richness for two (see Table 7). Thus, while various pronunciation variables still factored into listener judgments of comprehensibility, listeners this time also considered lexicogrammar when attempting to interpret meaning in speakers' utterances.

One implication of these findings is that linguistic impact on comprehensibility is a matter of degree, determined by the relative weighting of pronunciation and lexicogrammar variables, with pronunciation aspects of L2 speech likely having a consistent (and substantial) contribution to comprehensibility in all tasks, regardless of their complexity, and the additive contribution of lexicogrammar variables contingent on task complexity. Indeed, pronunciation and fluency variables mattered for comprehensibility in both tasks, and the role of these variables

in fact *increased* in the TOEFL task, compared to the IELTS task, as shown by the results of the PCA and follow-up regression analyses. The same increase was evident in a greater number of significant associations between comprehensibility and pronunciation/fluency variables in the TOEFL task, compared to the IELTS task (10 vs. 8), as shown in Tables 6 and 7. What this implies, then, is that pronunciation and fluency characteristics of L2 speech represent a substantial challenge to comprehensibility across a range of tasks, whereas the contribution of the lexis, grammar, and discourse content in L2 learners' speech likely grows as complexity increases.

<B>Task Complexity and L1 Effects

Although investigating L1 effects on comprehensibility was not the primary goal of this study (see Crowther et al., in press, for a detailed report), examining four L1-based groups separately, as opposed to pooling speakers together, proved advantageous. From a theoretical standpoint, nearly all conceptual frameworks of L2 speech learning predict L1-specific influences on production (e.g., Eckman, 2004; Escudero & Boersma, 2004; Flege, 2003). And from a practical perspective, teachers are acutely aware of L2 speakers' pronunciation difficulties traceable to their L1, as reflected in some pedagogical materials (e.g., Swan & Smith, 2001). Unsurprisingly, the current dataset revealed L1-specific influences on the relationship between comprehensibility and linguistic dimensions in L2 speech. For instance, in the TOEFL task, Romance speakers demonstrated the pattern of significant associations which was previously obtained for French speakers (Saito et al., under review; Trofimovich & Isaacs, 2012), with segmental errors, word stress/rhythm, fluency, and lexis associated with comprehensibility. In the same task, Chinese speakers showed a strong association between segmental accuracy and comprehensibility, which likely stems from the challenge that segmental production poses to these speakers, leading to more substitutions and errors of syllable structure (Anderson–Hsieh, Johnson, & Koehler, 1992; Rau et al., 2009). For Hindi/Urdu speakers in the same task, a strong association obtained for intonation may be related to how intonation in Hindi/Urdu is used to indicate stress through an increase in pitch (Shackle, 2001), which may have been distracting for listeners. Because this group was primarily composed of male speakers (14/15), it is important to determine whether similar findings would be found in a more gender-balanced group. The Farsi group was unique in that their comprehensibility was associated with lexicogrammar variables even in the simpler IELTS task. It is possible that these speakers relied on rather complex vocabulary and structure even in this task, which increased listeners' sensitivity to these variables. In fact, an informal analysis of written transcripts showed that Farsi speakers, compared to other groups, used academic words (*generate*, *communicate*, *socialize*) as well as relatively complex structures (participial and infinitival complements, conditional clauses) in the task. In essence, both L1-specific and individual, speaker-related factors might influence what listeners attend to in their perception of comprehensibility.

The final analysis of this study speaks directly to the relationship between task complexity and L1 effects. In particular, in the IELTS task, membership of the 60 speakers in their respective L1 groups could be predicted through three individual factors: lexical and discourse richness, rhythm and speech rate, or segmental and word stress accuracy. In essence, better performance in any (or all) of these factors could discriminate among the speakers as members of their L1 groups, implying that, for speakers of different L1s, comprehensible speech is linked to combinations of different (L1-specific) factors. In contrast, to successfully discriminate among the speakers in the TOEFL task, a single function was required, one that embraced all 10 linguistic categories. This finding suggests an intriguing possibility that

increased cognitive task complexity might be associated with *diminished* L1 influences on comprehensibility. Put differently, increased cognitive demands of complex tasks require speakers to attend to multiple linguistic dimensions at once to get their meaning across in a comprehensible manner. This ensures that, in complex tasks such as the TOEFL task in this study, comprehensibility for speakers from different linguistic backgrounds is no longer linked to a few linguistic variables specific to their L1s. Clearly, this possibility needs to be investigated further.

<A>IMPLICATIONS AND CONCLUSIONS

The findings of this study overall support previous research into L2 comprehensibility. Comprehensibility appears to be linked to a wide range of linguistic variables (Kang et al., 2010; Saito et al., in press; Trofimovich & Isaacs, 2012) whose strength varies as a function of speakers' L1 (Crowther et al., in press). Assuming (as shown here) that these linguistic variables also depend on a speaking task, with cognitively more demanding tasks drawing on a wider range of variables, compared to simpler tasks, these findings have several promising practical implications. For instance, they support what many language teachers already know from experience, namely, that teaching pronunciation targets beyond individual sounds, such as syllable structure, word stress, and fluency phenomena, is worth targeting in instruction (Foote et al., 2011). These results also imply that teachers should continue raising learners' awareness of how fluency, grammar, and lexical knowledge affect listener understanding. Because these linguistic factors are, to an extent, dependent on task, teachers should engage students in diverse speaking activities, with the goal of ensuring ongoing communicative success in a variety of contexts, and particularly those that resemble real-world domains, including assessment situations, in which speakers will need to perform.

Another practical consideration concerns integrating instruction targeting comprehensibility with assessment in L2 classrooms. As argued by Isaacs and Trofimovich (2012), language practitioners would benefit from a pedagogically-oriented assessment instrument targeting comprehensibility to guide them in instruction and assessment, which would be consistent with a focus on comprehensible speech over accent reduction (Derwing & Munro, 2009; Levis, 2005). Based on the current findings, however, such an instrument would need to be validated across several tasks, so that assessment rubrics could be adapted for particular task difficulty (e.g., in terms of cognitive demands). Or different comprehensibility scales could be empirically derived for different task types that learners are likely to engage with in the real world (Upshur & Turner, 1999). What is important in the future, then, is for researchers, assessment specialists, and teachers to embrace a key finding of this study that confirms results in other areas of L2 research, including speaking proficiency (e.g., Brown, Iwashita, & McNamara, 2005), namely, that there are different paths by which speakers can achieve comprehensible L2 speech, and that the artifact of the task has some bearing on speaker output and the quality of the speaker or test-taker performance in terms of linguistic performance.

ACKNOWLEDGMENTS

This study was funded by grants from the Social Sciences and Humanities Research Council of Canada (SSHRC) and Fonds de recherche sur la société et la culture (FRQSC) awarded to the second author, a Marie Curie Career Integration Grant (European Commission) awarded to the third author, and by the Grant-in-Aid for Scientific Research in Japan (No. 26770202) awarded

to the fourth author. We are grateful to participants, Ze Shan Yao for his technical help, and to anonymous MLJ reviewers for their helpful input and feedback on the content of this manuscript.

NOTES

1. An anonymous reviewer raised the question of whether linguistically trained raters familiar with accented speech could serve as a methodological benchmark for future research, suggesting either bilingual or multicompetent raters as another possibility. Because previous research has shown that both accent familiarity among native-speaking raters (e.g., Winke, Gass, & Myford, 2013) and matching/mismatching linguistic backgrounds of non-native speakers and listeners (e.g., Major, Fitzmaurice, Bunta, & Balasubramanian, 2002) can influence speech ratings, future research should consider task effects on L2 comprehensibility as a function of rater status (e.g., native speaker vs. L2 user).

2. Raters identified grammatical understanding as being the most understandable (M = 8.70, SD = .48) and intonation as least understandable (M = 7.80, SD = .92). For ease of use, lexical appropriateness was rated highest (M = 8.30, SD = .95) and rhythm as lowest (M = 7.20, SD = 1.62).

3. Further information about the development and validation of the linguistic rating scale used in the current study can be found in Saito et al. (in review).

4. A principal component analysis investigates which linear components (referred to here as factors) exist within a data set and how particular variables may contribute to these components. The oblimin rotation used here is an oblique rotation applied when there are theoretical grounds to believe that different variables of interest may correlate (Field, 2009), which was likely the case here with various linguistic dimensions of L2 speech.

5. The Kaiser–Meyer–Oklin test and Bartlett's test of sphericity are used to test the assumption of factorability for principal component analysis. These tests ensure that an appropriate level of correlations exists between variables to effectively run such an analysis.

6. Discriminant analysis is an approach that uses linear combinations of dependent variables (in this case, 10 individual linguistic scores) to allow for separation or discrimination between participant groups (Field, 2009), which in this case corresponds to predicting each speaker's membership in his or her L1 group.

REFERENCES

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning, 42*, 529–555.

Authors (xxxx).

Bongaerts, T. (1999). Ultimate attainment in L2 pronunciation: The case of very advanced late learners of Dutch as a second language. In D. P. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 133–59). Mahwah, NJ: Lawrence Erlbaum.

Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English for academic purposes speaking tasks. *TOEFL Monograph, 29*. Princeton, NJ: Educational Testing Service.

Chalhoub–Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System, 28*, 523–539.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*.
Crowther, D et al. in press
Derwing, T. M. & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching, 42*, 476–490.
Derwing, T. M. (2003). What do ESL students say about their accents? *The Canadian Modern Language Review, 59*, 476–490.
Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. *Studies in Second Language Acquisition, 20*, 1–16.
Derwing, T. M., Munro, M. J., & Thomson, R. I. (2008). A longitudinal study of ESL learners' fluency and comprehensibility development. *Applied Linguistics, 29*, 359–380.
Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*, 655–679.
Dickerson, L. (1975). The learner's interlanguage as a system of variable rules. *TESOL Quarterly, 9*, 401–407.
Duanmu, S. (2007). *The phonology of standard Chinese: The phonology of the world's languages.* Oxford: Oxford University Press.
Eckman, F. (2004). From phonemic differences to constraint rankings: Research on second language phonology. *Studies in Second Language Acquisition, 26*, 513–549.
Educational testing service. (2006). *The official guide to the new TOEFL iBT*. North America: McGraw Hill.
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, MI: The University of Michigan Press.
Ellis, R. (2001). Non-reciprocal tasks, comprehension and second language acquisition. In M. Bygate, P. Skehan, & M.Swain (Eds.), *Researching pedagogical tasks: Second language learning, teaching and testing* (pp. 49–74). Essex, UK: Pearson Education.
Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition, 26*, 551–585.
Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*, 313–326.
Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.
Field, J. (2009). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*, 399–423.
Flege, J. (2003). Assessing constraints on second-language segmental production and perception. In A. Meyer & N. Schiller (Eds.), Phonetics and phonology in language comprehension and production: Differences and similarities (pp. 319–355). Berlin: Mouton de Gruyter.
Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*, 1–26.
Hahn, L. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*, 201–223.
Harding, L. (2013). Pronunciation assessment. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Malden, MA: Wiley–Blackwell online.
DOI: 10.1002/9781405198431.wbeal0966
Hilton, H. (2008). The link between vocabulary knowledge and spoken L2 fluency. *The Language Learning Journal, 36*, 153–166.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist*. London: SAGE.
IELTS. (2009). *Official IELTS practice materials*. IELTS International: Los Angeles.
Isaacs, T., & Trofimovich, P. (2011). *International students at Canadian universities: Validating a pedagogically-oriented pronunciation scale*. Unpublished corpus of second language speech.
Issacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition, 34*, 475-505.
Jakeman, V. & McDowell, C. (2008). *New insight into IELTS: Student's book with answers*. Cambridge: Cambridge University Press.
Jenkins, J. (2013). English as a Lingua Franca in the international university: The politics of academic English language policy. New York: Routledge.
Jun, S.–A. (2005). Prosodic typology. In S.–A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing* (pp. 430–458). Oxford: Oxford University Press.
Kang, O., Rubin, D., & Pickering, L. (2010). Supresegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal, 94*, 554–566.
Kormos, J. (2006). *Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.
Labov, W. (1966). *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.
Larson–Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.
Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly, 39*, 369–377.
Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of language acquisition: Second language acquisition* (pp. 413–468). New York: Academic Press.
MacKay, A. & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. MacKay (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407–452). Oxford: Oxford University Press.
MacKay, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition, 22*, 471–497.
MacKay, I. R. A., Flege, J. E., & Imai, S. (2006). Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics, 27*, 157–183.
Major, R. C. (2001). *Foreign accent: The ontogeny and phylogeny of second language phonology*. Mahwah, NJ: Lawrence Erlbaum.
Major, R. C., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*, 173–190.
Martin, J. R., & Rose, D. (2003). *Working with discourse: Meaning beyond the clause*. New York: Continuum.
Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age, motivation, and instruction. *Studies in Second Language Acquisition, 21*, 81–108.

Munro, M. J. (2003). A primer on accent discrimination in the Canadian context. *TESL Canada Journal, 20*, 38–51.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 49*, 285–310.

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

Munro, M. J., & Derwing, T. M. (2011). Research timeline: The foundations of accent and intelligibility in pronunciation research. *Language Teaching, 44*, 316–327.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental pragmatics* (pp. 43–72). New York: Academic Press.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly, 6*, 113–125.

Rau, D. V., Chang, H.–H. A., & Tarone, E. E. (2009). Think or sink: Chinese learners' acquisition of the English voiceless interdental fricative. *Language Learning, 59*, 581–621.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*, 27–57.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching, 43*, 1–32.

Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. del Pilar García Mayo (Ed.). *Investigating task in formal language learning* (pp. 7–26). Clevedon, UK: Multilingual Matters.

Saito, K., Trofimovich, P., & Isaacs, T. (in press). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*.

Saito, K., Trofimovich, P., & Isaacs, T. (under review). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. Is this "under review"??

Segalowitz, N. (2010). *The cognitive bases of second language fluency*. New York: Routledge.

Shackle, C. (2001). Speakers of South Asian languages. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (2nd ed., pp. 310–324). Cambridge: Cambridge University Press.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*, 510–532.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185–211.

Swan, M., & Smith, B. (Eds.). (2001). *Learner English: A teacher's guide to interference* (2nd ed.). Cambridge: Cambridge University Press.

Tarone, E. (1983). On the variability of interlanguage systems. *Applied Linguistics, 4*, 142–164.

Tokumoto, M., & Shibata, M. (2011). Asian varieties of English: Attitudes towards pronunciation. *World Englishes, 30*, 392–408.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*, 905–916.

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing, 16*, 82–111.

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

Wilson, L., & Wilson, M. (2001). Farsi speakers. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems* (2nd ed., pp. 310–324). Cambridge: Cambridge University Press.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*, 231–252.

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.