Flawed self-assessment: Investigating self- and other-perception of second language speech

PAVEL TROFIMOVICH
Concordia University
Department of Education
1445 de Maisonneuve Blvd., West
Montreal, Quebec
Canada H3G 1M8
E-mail: 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

SARA KENNEDY
Concordia University
Department of Education
1445 de Maisonneuve Blvd., West
Montreal, Quebec
Canada H3G 1M8
E-mail: sara.kennedy@education.concordia.ca

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

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

Manuscript received: 29-Jun-2014
Revision received: 27-Oct-2014
Final acceptance: 14-Nov-2014

Acknowledgements

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 first, second, and third authors, a Marie Curie Career Integration Grant (European Commission) awarded to the second 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 BLC reviewers for their helpful input and feedback on the content of this manuscript.

Highlights

- L2 speakers (*N* = 134) self-assessed their accenteness and comprehensibility.
- Results revealed inaccurate self-assessment, compared to listener-rated measures.
- Discrepancies linked to phonology and fluency, not lexis, grammar, or discourse.
- Results imply ways of helping L2 speakers align self-assessment with performance.

Abstract

This study targeted the relationship between self- and other-assessment of accentedness and comprehensibility in second language (L2) speech, extending prior social and cognitive research documenting weak or non-existing links between people's self-assessment and objective measures of performance. Results of two experiments (*N* = 134) revealed mostly inaccurate self-assessment, with speakers at the low end of the accentedness and comprehensibility scales overestimating their performance and speakers at the high end of each scale underestimating it. For both accent and comprehensibility, discrepancies in self- versus other-assessment were associated with listener-rated measures of phonological accuracy and temporal fluency but not with listener-rated measures of lexical appropriateness and richness, grammatical accuracy and complexity, or discourse structure. Findings suggest that inaccurate self-assessment is linked to the inherent complexity of L2 perception and production as cognitive skills and point to several ways of helping L2 speakers align or calibrate their self-assessment with their actual performance.

People are famously poor at judging their own ability, engaging in such behaviours as "errors of omission", "flawed self-assessment", and "faulty self-awareness" (Carter & Dunning, 2008). It is an established finding that people's self-assessment and objective measures of performance relate poorly if at all. Meager or non-existing links between self- and other-assessments seem to be evident in all aspects of human behaviour, from sports, to health, to education, to the workplace (Dunning, Heath, & Suls, 2004; Falchikov & Boud, 1989; Harris & Schaubroek, 1988). In fact, across different skill sets, mean correlations between expected and observed performance fluctuate around .29, with correlations in sports being the highest and in the social domain, such as interpersonal skills, the lowest (Mabe & West, 1982).

The relationship between self- and other-assessment is important. At the conceptual level, it reflects one aspect of human metacognition, namely, the ability to evaluate both one's own and other people's competence across domains (Klin, Guzman, & Levine, 1997; Lichtenstein & Fischoff, 1977). And at the practical level, being able to adequately judge one's performance is linked to real-world decision-making, with inaccurate self-assessments leading people to engage in potentially dangerous, skill-inappropriate behaviours (e.g., poor drivers attempting to drive in adverse weather conditions, executives taking untenable financial risks) or to abstain from beneficial experiences (e.g., students dropping skill-appropriate courses).

Although the body of literature on self-assessment is wide-ranging, it is far from complete. One notable absence is research targeting second language (L2) pronunciation, which refers here to the linguistic characteristics underlying listener-based global constructs such as accentedness (nativelikeness) and comprehensibility (ease of understanding) in L2 speech.[1] Apart from a handful studies (e.g., Dlaska & Krekeler, 2008), it is unknown whether L2 speakers might misjudge their pronunciation (i.e., with respect to accentedness or comprehensibility), relative to the judgment of others, or what linguistic variables might underlie such assessment. Understanding the link between self- and other-assessment of L2 pronunciation is essential, given the importance of international trade and education, combined with the ever-growing interest in global popular culture and social media. These factors underscore the need for speakers to achieve communicative success in multiple languages, especially in pronunciation. This is because listeners differ in their tolerance for foreign accent (Moyer, 2013) and because comprehensibility is essential for efficiently communicating with an interlocutor (Isaacs & Trofimovich, 2012). If L2 speakers hold distorted views of their own pronunciation abilities, attaining communicative success will be problematic at best. Therefore, the goal of this study was to determine if there is a gap between self- and other-assessment of L2 speakers' pronunciation skills and to examine what linguistic factors might underlie this mismatch.

Dunning-Kruger Effect

When people compare their own ability and performance to those of others, a common finding is that more than half will judge themselves to be better than average, and those with poorer ability will be more likely to overestimate it than those with better skills (Carter & Dunning, 2008; Dunning et al., 2004). That poor performers tend to misjudge their ability has come to be known as the Dunning-Kruger effect (Dunning, Johnson, Ehrlinger, & Kruger, 2003; Kruger & Dunning, 1999). This effect is chiefly attributed to the information deficit faced by people at a lower level of skill. Poor performers suffer from a double "curse": not only are they poor at a particular skill, such as math or driving, but inadequate skill also prevents them from accurately evaluating their own competence. Put simply, compared to skilled performers, unskilled ones tend to overrate themselves because they are unaware of their incompetence. For instance, college students scoring in the bottom quartile on a psychology exam overestimate their score by about 30%, compared to the students in the top quartile whose self-assessment is more closely calibrated with their scores (Kruger & Dunning, 1999). This effect emerges even when people are promised money for accurate self-assessment, when they receive feedback about their performance, or when poor performers are given the chance to familiarize themselves with the performance of peers (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Mattern, Burrus, & Shaw, 2010; Simons, 2013; Sinkavich, 1995).

People's self-assessment also depends on a specific domain in which they evaluate themselves, with self-assessment in ill-defined, fuzzy, or highly subjective domains, such as intelligence, sophistication, or idealism, being more inflated than in more specific domains like neatness or punctuality (Burson, Larrick, & Klayman, 2006; Dunning, Meyerowitz, & Holzberg, 1989; Hayes & Dunning, 1997). Inaccurate self-assessment is also thought to be due to missing, ambiguous, or biased feedback which does not allow people to get accurate impressions of themselves. For instance, positive feedback is often rare (Dunning, 2005), such that people frequently do not know that they are performing well. Negative feedback is commonly worded to protect people's feelings and egos, with the consequence that it provides little concrete evidence to shape future behavior (Tesser & Rosen, 1975).

And it is not only poor performers that are prone to inaccurate self-assessment. Unlike poor performers, who overestimate their expertise, top performers tend to be overly modest, underestimating their ability relative to the perception of others (Ehrlinger et al., 2008; Kruger & Dunning, 1999). The source of this error in self-judgment is likely not the lack of skill but rather attribution of success to others (Fussell & Krauss, 1992), although top performers' negative self-assessment appears to be easy to correct by showing them the performance of peers (Hodges, Regehr, & Martin, 2001; Kruger & Dunning, 1999). Nonetheless, as with overinflated self-assessment by poor performers, top performers' underestimated self-assessment may have behavioral consequences. For instance, people who feel that they perform worse than they actually do on a quiz of scientific reasoning tend to decline a future invitation to participate in a science competition (Ehrlinger & Dunning, 2003). In sum, for both unskilled and skilled performers, perceptions – rather than reality – may determine decision-making.

Self-Assessment in L2 Pronunciation Research

The construct of self-assessment has had a substantial presence in the field of L2 learning and assessment (see Upshur, 1975), with research focusing on self-assessment in test validation (e.g., Weigle, 2010), as an awareness-raising tool (e.g., Glover, 2011), as predictor of learners' in-class participation (e.g., de Saint Léger, 2009), and a key element in learner-centered learning and teaching (e.g., Little, 2005). In addition, self-assessment has been seen as integral to the L2 proficiency construct, with self-assessment used as diagnostic feedback on reading, writing, and listening to complement test-takers' actual test performance, for instance, in the DIALANG diagnostic tests assessing proficiency in 14 European languages (Alderson, 2005). In L2 literature, findings in line with the Dunning-Kruger effect have also been reported, with low correlations between self- and other-assessed skills (e.g., Blanche & Merino, 1989; Brantmeier, Vanderplank, & Strube, 2012; Ross, 1998). And language assessment specialists have frequently noted, again consistent with the Dunning-Kruger effect, that learners at a lower level of ability overestimate their L2 performance (Davidson & Henning, 1985; Janssen van Dieten, 1989) while higher-ability learners underestimate it (Heilenman, 1990).

However, one L2 skill set that is mostly missing from this literature is pronunciation. Pronunciation, which encompasses dimensions associated with linguistic attributes of spoken language (e.g., prosody, segmental accuracy), is arguably one of the hardest skills to acquire.

Adult L2 speakers rarely sound nativelike, with accented L2 speech generally seen normal and often unavoidable, even for early bilinguals (Flege, Munro, & MacKay, 1995). Despite this, many learners view the linguistic ability of a native speaker, characterized by near-native accent, as their ideal ultimate learning goal (Tokumoto & Shibata, 2011). At the same time, many learners, while perhaps aware that they have difficulties with pronunciation, might be unable to identify their specific linguistic problems because learners lack diagnostic abilities or metalinguistic knowledge to articulate them, particularly at lower ability levels (Derwing & Rossiter, 2002). Pronunciation is rarely targeted in communicative L2 classrooms. For instance, in a classroom-based study by Foote, Trofimovich, Collins, and Soler Urzúa (2013), a focus on pronunciation accounted for about 10% of all language-related episodes. Similarly, teachers may also provide only limited types of feedback targeting pronunciation, mainly in the form of implicit error correction such as recasting (Lyster, 2001), with the consequence that learners may not recognize feedback as such, particularly at lower proficiency levels (Ammar & Spada, 2006), or may not benefit from feedback due to learners' lack of experience in pronunciation training and relevant linguistic knowledge (Saito, 2013). Finally, pronunciation difficulties are highly context-dependent, with particular difficulties linked to speakers' linguistic backgrounds and also to listener characteristics, such as being native or non-native interlocutors (Smiljanic & Bradlow, 2009). L2 pronunciation thus represents an ill-defined and highly complex skill with often missing or ambiguous feedback. As such, L2 speakers would likely be susceptible to faulty self-assessment.

However, with respect to L2 pronunciation, there is limited evidence about discrepancies in self- and other-assessment and the linguistic dimensions which underlie these discrepancies. For example, Yule and his colleagues examined the accuracy of L2 English learners' segmental perception (e.g., *cloud* vs. *crowd*) and their confidence ratings about their accuracy at two times separated by seven weeks (Yule, Damico, & Hoffman, 1987; Yule, Hoffman, & Damico, 1987). Little relationship was observed between learners' self-confidence of perception accuracy and their actual accuracy, with some learners improving only in self-confidence but others only in perception accuracy. More recently, focusing on the production of German vowels and consonants, Dlaska and Krekeler (2008) asked advanced learners of German to compare their own production of vowels and consonants to native-speaker models. Whereas the learners and two trained listeners mostly agreed on which sounds were produced accurately, the learners were able to identify only 44% of their vowel and consonant errors, revealing a gap between self- and other-assessment of segmental production. In sum, existing evidence suggests that L2 speakers' self-assessment is misaligned with their actual performance, yet it is unclear to what extent L2 speakers are over- or under-confident in their self-judgment or which aspects of speech, apart from segmentals, are subject to distorted self-assessment.

The Current Study

With the goal of clarifying the relationship between self- and other-assessment of L2 pronunciation, the current study targeted L2 speakers from multiple language backgrounds whose speech, recorded in an extemporaneous speaking task, was rated by the speakers themselves and also by native-speaking listeners for two constructs (accent, comprehensibility). Of primary interest was a potential discrepancy between speakers' own and native-speaking listeners' assessment (Dunning-Kruger effect). However, besides exploring this effect in the domain of L2 pronunciation, which (as argued previously) represents a complex and ill-defined skill with frequently missing or ambiguous feedback, this study also aimed to examine which linguistic aspects of L2 speech might be linked to inaccuracies in L2 speakers' self-assessment.

These aspects included a wide range of linguistic factors that feed into overall conceptions of L2 performance as reflected in rating scales, including segmental and suprasegmental accuracy, temporal fluency, lexical appropriateness and richness, grammatical accuracy and complexity, as well as discourse structure.

To address these goals, this study focused on two broad constructs of L2 pronunciation: accentedness and comprehensibility. Accentedness (a measure of linguistic nativelikeness) refers to listeners' perceptions of how closely speakers can approximate speech patterns of the target-language community, while comprehensibility (a broad measure of speakers' communicative effectiveness) is defined as listeners' perceptions of how easily they can understand L2 speech. Accentedness and comprehensibility are overlapping yet distinct constructs, as illustrated by the finding that even some heavily accented L2 speech can be highly comprehensible (Derwing & Munro, 2009). These two dimensions are particularly relevant to investigating self-assessment because they represent two common learning and teaching goals (Levis, 2005), namely, a focus on nativelikeness (accent reduction) and a focus on understanding (comprehensibility).

This study thus investigated the relationship between self- and other-assessment of L2 speakers' accent and comprehensibility through two experiments, addressing two questions. The question targeted in Experiment 1 was "Do L2 speakers demonstrate a discrepancy between their own and native-speaking listeners' assessment of accentedness and comprehensibility in L2 speech?" The question targeted in Experiment 2 was "Which linguistic characteristics of L2 speech (including segmental and suprasegmental accuracy, temporal fluency, lexical appropriateness and richness, grammatical accuracy and complexity, and discourse structure) are most susceptible to discrepancies between self- and other-assessment of accentedness and comprehensibility in L2 speech?” The overall goal of both experiments was to expand the study of self-assessment behavior in the field of L2 speech learning, to better understand the linguistic dimensions that feed into L2 speakers’ awareness of accentedness and comprehensibility.

Experiment 1

Experiment 1 targeted the relationship between self- and other-assessment of L2 speech for a large sample of L2 speakers (*n* = 134) from multiple language backgrounds, with the goal of documenting how closely L2 speakers’ judgments of accent and comprehensibility relate to native-speaking listeners’ assessment of the same constructs. Consistent with previous research in the domains of social and cognitive psychology (e.g., Carter & Dunning, 2008) and L2 learning and assessment (e.g., Dlaska & Krekeler, 2008), it was predicted that L2 speakers would show discrepancies in their self-assessment, relative to the rating by native-speaking listeners, with speakers judged as most accented and least comprehensible overestimating their ability and speakers at the opposite end underestimating their ability.

Participants

The participants in Experiment 1 were 134 speakers (31 female, 103 male) with a mean age of 23.9 years (*SD* = 3.2) from an unpublished corpus of L2 speech by speakers from 19 language backgrounds completing five tasks (Isaacs & Trofimovich, 2011). The language groups represented in the corpus included speakers of Farsi (33), Chinese (14), Telugu (13), Hindi (12), Bengali (9), Tamil, French (8 each), Punjabi, Arabic (7 each), Spanish (7), Gujarati (4), Urdu (3), Greek, Marathi (2 each), as well as Akan, Kannada, Kinyarwanda, Malayalam, and Portuguese (1 each). The speakers, who were international students in undergraduate (25) and graduate (109) programs at an English-medium Canadian university, had studied English on average 21.5 years (*SD* = 5.7), primarily through formal instruction in primary, secondary, and university-level settings. They had arrived in Canada to pursue studies at a mean age of 23.3 years (*SD* = 3.8) and participated in the experiment during the first term of university studies, which usually started soon after their arrival in Canada. All speakers had recently taken either TOEFL iBT or IELTS tests, which are high-stakes instruments that were used to assess the participants' ability to pursue university studies (see Chalhoub-Deville & Turner, 2000). The participants' mean overall scores were 88.8 (*SD* = 9.6) for TOEFL iBT and 6.9 (*SD* = .6) for IELTS. The speakers self-rated their English ability at a mean of 6.6 (*SD* = 1.1) in speaking and 7.3 (*SD* = 1.3) in listening using 9-point Likert-type scales (1 = *extremely poor*, 9 = *extremely fluent*). Using 0-100% scales (0% = *never*, 100% = *all the time*), they also estimated their daily use of English at 61.6% (*SD* = 22.9) in speaking and at 70.1% (*SD* = 20.5) in listening. The speakers indicated that they used English at the university on average 76.6% daily (*SD* = 23.5).

Materials and Procedure

As part of the original corpus, each speaker completed five speaking tasks administered in a randomized order, but only the picture story task, one of the most common tasks used to elicit L2 speech (e.g., Trofimovich & Isaacs, 2012; Derwing, Rossiter, Munro, & Thomson, 2004), was chosen for analysis in this study. The task consisted of an eight-frame colored picture narrative depicting a man and a woman who collided with each other on a busy street corner, accidentally switched their identical suitcases, and finally realized their error after arriving in their respective destinations (Derwing et al., 2004). The speakers were first shown the picture sequence and were instructed to describe the story by explaining what happened in each image, with no time limit imposed. The narratives were recorded directly onto a computer and stored as digital audio files. After completing the task, the speakers were asked to use a 9-point scale to indicate how well they performed it (1 = *very poorly*, 9 = *very well*) and to estimate overall task difficulty (1 = *very easy*, 9 = *very difficult*), in that order. They were subsequently instructed to self-rate their accent, which refers to the extent of native language (L1) influences in their speech (1 = *heavily accented*, 9 = *not accented at all*), and comprehensibility, which denotes presumed listener effort in understanding their speech during the task (1 = *hard to understand*, 9 = *easy to understand*).

The audio files were subsequently normalized by matching peak amplitude across files, then presented to listeners. The listeners were three native English speakers (all females), with a mean age of 41.7 (*SD* = 8.1) and 10 years (*SD* = 0) of L2 teaching experience. They had been educated entirely in English, holding advanced degrees in applied linguistics or language teaching and completing at least one course on applied phonetics and pronunciation teaching. They reported using English daily on average 86.7% of the time (*SD* = 5.8) in speaking and 90% of the time (*SD* = 0) in listening, and reported extensive exposure to L2 English and some proficiency in another language (French, German, Japanese, Swahili). The listeners individually evaluated the 134 audio files, which were on average 68.3 s long (*SD* = 30.3), rating each file for the same two dimensions, namely, accent (1 = *heavily accented*, 9 = *not accented at all*) and comprehensibility (1 = *hard to understand*, 9 = *easy to understand*). The listeners, who used a personal computer for audio playback, were first given definitions of each rated construct and invited to discuss any questions. They then received a rating booklet and rated three practice files. The listeners worked at their own pace, playing each consecutive file and recording their ratings in the booklet, with an unlimited number of replays permitted. Although they were not required to play the entire file to make decisions, all listeners listened to at least 20-30 s of speech in each recording, which is consistent with 15-30 s samples used to obtain listeners' impressionistic ratings of speech in prior research (e.g., Derwing et al., 2004).

Analysis

Cronbach's alpha, a measure of rater consistency, was computed first across the three listeners' ratings, separately for accent and comprehensibility. The obtained coefficients were reasonably high (Stemler & Tsai, 2008), particularly given the small sample size (.89 for accent and .79 for comprehensibility), exceeding the benchmark value of .70-.80 (Larson-Hall, 2010). Therefore, a single accent and comprehensibility score was derived for each speaker by averaging across the three listeners' judgments. The measure of primary interest was speakers' overconfidence scores in accent and comprehensibility. Overconfidence scores were derived by subtracting the mean rating for each speaker from the speaker's self-rating and then expressing the obtained numerical difference as a proportion on a 9-point scale. Numerical differences which were positive corresponded to speakers overestimating their accent (nativelikeness) or comprehensibility (ease of understanding), relative to native-speaking listeners, while negative numerical differences represented speakers underestimating their accent or comprehensibility. Values around zero indicated self-ratings that were calibrated or aligned with listener judgments. The extent of the difference between self- and other-ratings was illustrated by the magnitude of the proportion.

Results and Discussion

The first set of analyses examined the overall relationship between the L2 speakers' actual performance (as rated by native listeners) and their self-ratings. Pearson correlation tests (two-tailed) revealed no association between the speakers' actual and self-rated scores for accent, $r(132) = .06$, $p = .50$, and only a weak association for comprehensibility, $r(132) = .18$, $p = .03$, suggesting that the relationship between actual and self-rated performance was tenuous at best. However, consistent with the Dunning-Kruger effect, there were moderate-to-strong associations between speakers' overconfidence scores and their actual performance, both for accent, *r*(132) = –.67, *p* < .0001, and comprehensibility, *r*(132) = –.56, *p* < .0001. In both cases, the associations (illustrated in Figure 1) were negative, indicating that more accented and less comprehensible speech was associated with greater overconfidence scores. This suggests that the speakers who were rated by native-speaking listeners as most accented and least comprehensible were those whose self-ratings of accent and comprehensibility were most inflated, compared to listener ratings. Both associations were also comparable in their strength, as was shown by Fisher *r*-to-*z* transformations conducted to explore statistical differences in correlation coefficient strength, *Z* = 1.57, *p* = .12.

FIGURE 1

To further qualify the relationships shown in Figure 1, the overconfidence scores for the bottom and top thirds of the speakers (*n* = 45) were then compared, separately for accent and comprehensibility. For accent, the bottom third of the speakers was significantly more overconfident (*M* = .27 or +2.4 points on a 9-point scale) than the top third (*M* = –.19 or –1.7 points), *t*(88) = 19.39, *p* < .001, *Cohen's d* (effect size) = 5.40, who were underconfident. For comprehensibility, although overconfidence scores generally clustered around the upper range of the scale (see Figure 1), the bottom third of the speakers was again significantly more overconfident (*M* = .18 or +1.6 points on a 9-point scale) than the top third (*M* = –.23 or –2 points), *t*(88) = 19.13, *p* < .001, *d* = 10.52, who again underestimated their performance. To sum up, the speakers' self-ratings related little to their actual performance, as rated by native-speaking listeners. Instead, the speakers tended to either over- or underestimate their performance. Put differently, speakers at the bottom of the accent and comprehensibility scale overestimated their performance while speakers at the top of each scale underestimated it.

The next analysis examined how various speaker characteristics related to their overconfidence scores. These characteristics included background profiles collected in a participant questionnaire, as well as the speakers' self-ratings of task difficulty and task performance success. With respect to interlocutor background characteristics, such as age, amount of prior English study, self-rated speaking and listening ability, age of first exposure to English, amount of daily English use, or TOEFL and IELTS test scores or listening and speaking subscores, there were no significant associations between any of these characteristics and overconfidence scores, $r < .15$, $p > .07$. Speakers' overconfidence scores also showed no significant associations with their own ratings of how well they performed each task and how they estimated overall task difficulty, $r < .08$, $p > .37$. In sum, speakers' background characteristics or their perception of task success and task difficulty bore no obvious relationship to the extent of overconfidence they demonstrated in judging their own accent and comprehensibility.

The final analysis sought to ascertain that the obtained pattern of findings was not specific to the particular measure of overconfidence used, namely, a numerical difference between self- and listener-based ratings expressed as a proportion on a 9-point ordinal scale. In line with previous psychological research on self-assessment (e.g., Burson et al., 2006; Kruger & Dunning, 1999), rated accent and comprehensibility values were first rank-ordered and then expressed as percentile scores by subtracting listener-rated performance from speakers' own estimates to derive a percentile-based measure of overconfidence. The resulting overconfidence scores matched closely the original measure of overconfidence for both accent, $r(132) = .98$, $p < .0001$, and comprehensibility, $r(132) = .96$, $p < .0001$, and the pattern of findings obtained with this new measure was identical to the one reported previously. The relationship between percentile-based measures of the speakers' actual and self-rated performance is illustrated in Figure 2, which shows speakers' perceived percentile rankings (self-rating, shown by a dashed line) and their actual test performance (listener rating, shown by a solid line) plotted separately for four speaker groups based on listener-rated performance quartile (bottom to top 25%).

FIGURE 2

As Figure 2 shows, the speakers who were rated by listeners in the bottom 25% of the sample overestimated their performance (self-rating higher than listener rating), while the speakers rated by listeners in the top 25% underestimated it (self-rating lower than listener rating). In fact, self-rating and actual performance were aligned only for speakers performing around the 50th percentile in accent and comprehensibility, which roughly corresponded to a rating of 4.8 for accent and 6.7 for comprehensibility on a 9-point scale. To sum up, L2 speakers showed discrepancies in judgments of their accentedness and comprehensibility compared to judgments by native-speaking listeners, with speakers at the bottom of each scale overestimating their performance and speakers at the top of each scale underestimating it by a similar degree. Notably, none of the speakers' language background characteristics nor their ratings of task difficulty and task success could explain the extent of discrepancy between the speakers' self-rated and actual performance.

Experiment 2

As predicted, Experiment 1 revealed discrepancies in L2 speakers' self-assessment, relative to listeners' rating, with those speakers who were rated most accented and least comprehensible overestimating their ability and speakers who were rated at the higher end of the ability continuum underestimating their ability. However, the extent of L2 speakers' over- or under-confident behavior was unrelated to their language background characteristics or their ratings of task difficulty and task success. Therefore, Experiment 2 targeted the relationship between self- and other-assessment of L2 speech in more detail, for a smaller cohort of the original speakers (*n* = 56), by analyzing the speakers' speech for 10 linguistic categories from the linguistic categories of segmental and suprasegmental accuracy, fluency, lexical and grammatical appropriateness and richness, and discourse structure. The goal of Experiment 2 was therefore to determine which linguistic categories in L2 speech are most susceptible to discrepancies between self- and other-assessment of accentedness and comprehensibility. Based on limited previous research targeting segmental aspects of L2 speech perception and production (Dlaska & Krekeler, 2008; Yule et al., 1987), it was predicted that L2 speakers would be especially prone to misjudging their segmental production accuracy. However, given a strong link between accent and segmental and suprasegmental accuracy and fluency as well as between comprehensibility and lexis, grammar, and discourse (Crowther, Trofimovich, Saito, & Isaacs, 2014; Trofimovich & Isaacs, 2012), it was also expected that speakers' self-assessments might be linked to other aspects of L2 speech, besides segmentals.

Participants

The participants in Experiment 2 included 60 speakers drawn from the sample of 134 participants in Experiment 1. The speakers were selected based on their L1 background and represented the largest cohorts in the corpus, yielding groups of Farsi, Mandarin Chinese, Hindi/Urdu, and Romance speakers (*n* = 15). L1 background was used as a grouping variable to isolate possible L1 influences on speakers' self-assessment, since linguistic dimensions of L2 speech are known to be L1-specific (e.g., Flege, 2003). The Hindi/Urdu group combined the speakers of both languages because the key difference between them is script-based (King, 1994), and the Romance group included all speakers of French and Spanish, which share a common language family and a similar syllable-timed rhythm (Jun, 2005). However, the data for four speakers were lost due to error, leaving altogether 56 speakers, with a nearly-equal distribution of speakers per group: Farsi (*n* = 15), Chinese (*n* = 14), Hindi/Urdu (*n* = 14), and Romance (*n* = 13), with eight French and five Spanish speakers. The speakers were matched as closely as possible across the four groups based on background characteristics (shown in Table 1), with the exception of male/female ratio in the Hindi/Urdu group, which mirrored the gender makeup of these speakers in the university. According to one-way ANOVAs, there were no significant between-group differences in any of the scores shown in Table 1, *F*s < 2.65, *p* > .06, except age of arrival, *F*(3, 52) = 3.99, *p* = .012, η_p^2 = .19, with the Farsi group being slightly older than the Chinese (*p* = .03) and the Romance (*p* = .04) groups at the time of arrival in Canada.

TABLE 1

Materials and Procedure

The 56 speakers' audio recordings from the picture story task, along with written transcripts of each recording, were then presented to trained raters for linguistic coding using 10 rated categories targeting the dimensions of phonology, fluency, lexis, grammar, and discourse. The raters included 10 native English speakers (7 females, 3 males), with a mean age of 32.7 years (*SD* = 10.2) and an average of 6.6 years of L2 teaching experience (*SD* = 6.8). The raters were either recent graduates or current students in applied linguistics from the same university community as the speakers. They estimated using English a mean of 89% of the time (*SD* = 8.8) in speaking and 85% of the time (*SD* = 13.5) in listening, and reported extensive exposure to L2 English and some proficiency in another language (French, Spanish, Korean). The raters evaluated the 56 speakers' performance in the picture story task in two individual rating sessions (about 2 h each), conducted within three weeks. The first session was devoted to the coding of audio recordings for five phonology- and fluency-based categories, while the second session was dedicated to evaluating orthographic transcripts for five lexical, grammatical, and discourse categories (described below). Transcripts of recorded picture narratives (rather than audio recordings) were used in the second session in order to remove pronunciation and fluency as possible confounds in judgments of lexis, grammar, and discourse (Crossley, Salsbury, & McNamara, 2014).

In all rating sessions, the raters used computer-based scales developed by Saito, Trofimovich, and Isaacs (forthcoming). For each rated measure, the scale featured a 1000-point continuous slider, run through the MATLAB interface, with endpoints clearly marked with a frowning face on the left (rating of 0) and a smiley face on the right (rating of 1000). The slider was initially fixed in the middle (rating of 500), with no numerical labels or marked intervals shown. At the start of each session, the raters were trained on the relevant rated categories (see Appendix) and were shown how to use the scale. Depending on the session, they then listened to four supplementary audio recordings or viewed four supplementary transcripts and rated them for practice by using the relevant scales, with each rating discussed to determine if the raters understood each rated category as intended. All audio recordings were edited to remove all fillers and false starts at the beginning of the file and were shortened to include only the initial 30 s of speech, consistent with prior research using 20-60 s samples to evaluate speech (e.g., Derwing et al., 2004). To minimize phonology and fluency influences on rater judgments of lexis, grammar, and discourse (Crossley et al., 2014) and to avoid any transcriber influence (Ochs, 1979), all orthographic transcripts of the speakers' task performance were modified to remove hesitation markers (e.g., um, uh), spelling clues signaling phonology-specific errors (e.g., *three* pronounced as *tree* was spelled as “three”), and punctuation. All relevant scales (i.e., five categories for audio or transcript rating) were visible simultaneously, and the raters were allowed to adjust their judgments on all scales before proceeding to the next recording (or transcript). All audio recordings or transcripts (depending on the session) were presented to each rater in a unique randomized order, and the raters could replay each recording or reread each transcript as often as necessary.

Coded Linguistic Categories

The raters evaluated each audio recording for the following five segmental, suprasegmental, and temporal categories (see Appendix for training materials and onscreen labels):

1. Segmental errors (1 = “frequent”, 1000 = “infrequent or absent”), defined as errors in the articulation of individual sounds within a word (e.g., *dat* instead of *that*; *pin* instead of *pen*), as well as any sounds erroneously deleted from or inserted into words (e.g., *’ouse* instead of *house*; *supray* instead of *spray*).
2. Word stress errors (1 = “frequent”, 1000 = “infrequent or absent”), defined as errors in the placement of primary stress (e.g., *com-pu-TER* instead of *com-PU-ter*, where capitals designate stress) or the absence of discernible stress, such that all syllables receive equal prominence (e.g., *com-pu-ter*).
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., Will you be home tomorrow↑) or falling tones at the end of statements (e.g., Yeah, I’ll stay at home↓).
4. Rhythm (1 = “unnatural”, 1000 = “natural”), defined as the difference in stress (emphasis) between content and function (grammatical) words. For instance, in the

sentence “They RAN to the STORE”, the words “ran” and “store” are content words and therefore are stressed more than the words “they”, “to”, and “the”, 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.

The raters evaluated each orthographic transcript for the following five lexical, grammatical, and discourse categories (see Appendix for training materials and onscreen labels):

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., “A man and a woman bumped into each other on a walkside”).
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 girl arrived home her dog was happy she arrived home”, compared to “The girl arrived home to find her dog overjoyed at her return”).
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., “What you are doing?”), morphology (e.g., “She go to school every day”), and agreement (e.g., “I will stay there for five day”).
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 man wore a black hat and he enjoyed his coffee", compared to "The man who was wearing a black hat was enjoying his coffee").

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 narrative sounds developed and sophisticated.

Analysis

Cronbach's alpha coefficients, computed to determine inter-rater reliability, showed that the raters were fairly consistent, demonstrating reliability values that surpassed benchmark values of .70-.80 (Larson-Hall, 2010) for segmental and suprasegmental accuracy ($a_{\text{segmentals}}$ = .94; $a_{\text{word stress}}$ = .87; $a_{\text{intonation}}$ = .84; a_{rhythm} = .86), fluency ($a_{\text{speech rate}}$ = .90), vocabulary ($a_{\text{appropriateness}}$ = .80; $a_{richness}$ = .86), grammar ($a_{\text{acccuracy}}$ = .82; $a_{complexity}$ = .85), and discourse ($a_{richness}$ = .88). Considered sufficiently consistent, the scores were averaged across the 10 raters to obtain a single mean score per speaker for each rated category. The raters also appeared to have little difficulty understanding and applying the rated categories, as shown through their feedback at the end of each session. They had estimated the extent to which they understood the categories at 8.3 (*SD* = .5) on a 9-point scale (1 = *I did not understand at all*, 9 = *I understand this concept well*) and rated the degree to which they could comfortably and easily use them at 7.8 (*SD* = .9) on a similar scale (1 = *very difficult*, 9 = *very easy and comfortable*). Because the original measure of overconfidence from Experiment 1 – namely, an overconfidence score expressed as a proportion on a 9-point scale – shared a substantial amount of variance with a percentile-based measure (96% for accent, 92% for comprehensibility), it was decided to use this original measure as the dependent variable of interest in Experiment 2.

Results and Discussion

The first analysis was used for data reduction purposes to explore underlying patterns among the 10 rated linguistic categories in the speakers' speech. The goal was to uncover possible common dimensions across the 10 rated linguistic categories, so that these dimensions could be related to the degree of speaker overconfidence in judging their own accent and comprehensibility, relative to listener ratings. First, the scores from the 10 linguistic categories for the entire set of 56 speakers were submitted to an exploratory Principal Component Analysis (PCA) with Oblimin rotation to determine if the categories showed any underlying patterns based on their clustering. Although the sample size was relatively small, the Kaiser-Meyer-Oklin value was .86, exceeding the required .60 for sampling adequacy and indicating excellent factorability of the correlation matrix (Hutcheson & Sofroniou, 1999). Moreover, Bartlett's test of sphericity yielded a highly significant value, $\chi^2(45) = 653.92$, $p < .0001$, suggesting that the correlations between the categories were sufficient for PCA.[2] The PCA yielded two factors accounting for 83.6% of total variance (shown in Table 2). Factor 1, labeled Phonology, consisted of the four segmental/suprasegmental categories and speech rate. Factor 2, named Lexicogrammar, comprised all vocabulary, grammar, and discourse-level categories, plus speech rate. In sum, the 10 linguistic categories patterned along two dimensions (phonology and lexicogrammar), with speech rate common to both. That a fluency variable would be linked to both phonology and lexicogrammar is unsurprising because fluent speech reflects efficient processing at multiple levels, including those of phonological encoding and articulation as well as lexical retrieval and grammatical assembly (see Segalowitz, 2010).

TABLE 2

The next analysis targeted possible contributions of the PCA phonology and lexicogrammar factors to speakers' overconfidence. More specifically, the phonology and lexicogrammar PCA scores, computed through the Anderson-Rubin method of obtaining non-correlated factor scores (Field, 2009), were used as two predictor variables in a stepwise multiple regression analysis to examine the contribution of phonology and lexicogrammar to speakers' overconfidence scores, separately for accent and comprehensibility. The two regression models (*n* = 56) yielded nearly identical findings (summarized in Table 3), with the phonology factor emerging as the only significant predictor of overconfidence scores, accounting for 28% of variance for accent and 27% for comprehensibility.[3] Thus, as indicated by negative beta values (reflecting negative associations between variables) in Table 3, the extent of speakers' overconfidence was inversely related to their L2 phonology, with less nativelike segmental and suprasegmental accuracy and fluency linked to overestimated self-ratings of accent and comprehensibility, relative to listener ratings.

TABLE 3

This obtained relationship between speakers' overconfidence and their L2 speech is clarified further in Table 4, which shows individual Pearson correlations (two-tailed) between the speakers' overconfidence scores and their scores for the five linguistic categories subsumed under the PCA phonology factor. Each of the five categories accounted for 10-35% of shared variance in speakers' overconfidence in accent and 16-29% in their overconfidence in comprehensibility. In all cases, less accuracy in segmental production, in word stress, rhythm, and intonation, as well as slower speech rate were associated with overestimated self-ratings of accent and comprehensibility, relative to native listener ratings.

TABLE 4

The final analyses targeted between-group differences in overconfidence, on the assumption that speakers' self-assessment behaviours might be specific to their linguistic background. First, the overconfidence scores for the speakers in the four groups were submitted to one-way ANOVAs to determine if the extent of overconfidence varied as a function of the speakers' L1. These analyses yielded significant *F*-ratios for both accent, $F(3, 52) = 4.89$, $p = .005$, $\eta_p^2 = .22$, and comprehensibility, $F(3, 52) = 4.50$, $p = .007$, $\eta_p^2 = .21$. For accent, Bonferroni-corrected post-hoc tests revealed that the Chinese ($p = .039$) and the Hindi/Urdu ($p = .013$) speakers' overconfidence scores were significantly greater than those of the Romance speakers. For comprehensibility, the Chinese speakers' overconfidence scores were again significantly greater than those of the Romance ($p = .033$) and the Farsi speakers ($p = .021$). This pattern of findings is best illustrated in Table 5, which shows descriptive statistics for each group. For accent, the Chinese ($M = .12$) and Hindi/Urdu ($M = .15$) speakers overall tended to overestimate their performance, compared to the Farsi speakers ($M = –.04$), whose overconfidence scores were close to 0 (i.e., aligned with listener assessment), and the Romance speakers ($M = –.11$), who underestimated their performance. For comprehensibility, the Chinese speakers ($M = .09$) as a group overestimated their performance, compared to the Hindi/Urdu speakers ($M = .02$), whose assessment was aligned with native listener assessment, and the Farsi ($M = –.14$) and Romance ($M = –.14$) speakers, who both underestimated their performance. Given a strong inverse relationship between overconfidence scores and actual performance, as shown in Experiment 1, these findings imply that the groups that were more accented (i.e., Chinese, Hindi/Urdu) and less comprehensible (i.e., Chinese), as rated by native-speaking listeners, are those that significantly overestimated their performance, compared to the other groups.[4]

TABLE 5

Because the confidence scores for each group ranged widely between negative (underestimated performance) and positive (overestimated performance) values (see Table 5), it was then possible to examine which linguistic categories were associated with overconfidence, separately for each group. Table 6 shows the results of Pearson correlations (two-tailed) between overconfidence scores for each group and their scores for the five linguistic categories subsumed under the PCA phonology factor. The obtained pattern of significant correlations, taken together with the results of between-group comparisons illustrated in Table 5, suggests two broad conclusions. Firstly, for both accent and comprehensibility, the link between speakers' overconfidence and their segmental, suprasegmental, and fluency performance is strongest for groups with weaker actual performance (i.e., Chinese and Hindi/Urdu vs. Romance). For speakers rated lower in accent and comprehensibility (Chinese, Hindi/Urdu), segmental and suprasegmental accuracy and fluency likely feed into their inaccurate self-assessment. In contrast, for speakers at higher levels of accent and comprehensibility (Romance and Farsi, especially for comprehensibility), self-assessment is likely based on factors other than those targeted here.[5] And secondly, the specific segmental, suprasegmental, and fluency variables linked to overconfidence unsurprisingly appear to vary as a function of the speakers' L1 background, such that speakers' self-assessment is based on the variables that are likely most problematic for them, such as segmentals and intonation for Chinese speakers (Anderson-Hsieh, Johnson, & Koehler, 1992) or suprasegmentals (word stress, rhythm, intonation) for Hindi/Urdu speakers (Shackle, 2001).

TABLE 6

General Discussion

The two research questions of this study asked whether L2 speakers show discrepancies between their own and listeners' assessment of accentedness and comprehensibility in their speech and which linguistic characteristics of L2 speech are linked to such discrepancies. In line with previous research on the Dunning-Kruger effect (Carter & Dunning, 2008) and research in L2 assessment and learning (Davidson & Henning, 1985), the L2 speakers in this study showed mostly inaccurate self-assessment of how accented and comprehensible they sounded, relative to the ratings of native speakers. Consistent with prior findings, speakers at the low end of the accentedness and comprehensibility scales tended to overestimate their performance, while speakers at the high end of each scale underestimated it. In fact, only about a third of all L2 speakers were fully calibrated with listeners in their self-assessment (37/134 or 28% for accent and 51/134 or 38% for comprehensibility), placing themselves within ±10% of listener ratings. For both accent and comprehensibility, discrepancies in self- versus other-assessment were associated with several segmental and suprasegmental dimensions of L2 speech (segmental accuracy, word stress, rhythm, intonation, speech rate) but not with aspects of lexis, grammar, and discourse. Generally, the L1 groups that showed weaker performance (Chinese) tended to rate themselves more overconfidently than higher-scoring groups (Romance, Farsi), and the specific linguistic aspects of L2 speech feeding into inaccurate self-assessment appeared to be L1-specific, such as segmentals and intonation for Chinese speakers. These findings extend literature on self-assessment in social, academic, and professional domains (Dunning et al., 2004; Mabe & West, 1982) by showing that L2 speakers' rating of their accent and comprehensibility correspond only weakly to listeners' judgment. Also, given the argument that pronunciation is a complex skill for L2 speakers, such discrepancies are in line with prior research illustrating larger gaps between self- and other-assessment in more complex, as compared to simpler, skills and tasks (Burson et al., 2006; Hayes & Dunning, 1997).

Inaccurate self-assessment

A weak relationship between self- and other-assessment of L2 speech has interesting consequences for L2 speech learning, particularly within interactionist approaches to L2 development. Underlying interactionist approaches (e.g., Long, 1996) is the idea that specific aspects of interaction – referred to broadly as negotiation for meaning – ultimately lead L2 users to notice the discrepancy (i.e., the gap) between the target language and their own understanding of it (Long, 1991; Schmidt, 2001), which in turn facilitates language development (for review, see Mackey & Goo, 2011). Put simply, interaction-driven learning requires L2 users to notice similarities or differences between their own linguistic performance and the language produced by their interlocutors.

However, if L2 speakers consistently misjudge their performance relative to more objective measures, then they might have difficulty noticing the important ways in which their own production differs from targetlike language, particularly at low ability levels (Ammar & Spada, 2006). Consistent with the Dunning-Kruger effect, L2 speakers at the lower end of the speaking ability spectrum might be overconfident in their self-assessment, making it harder for them to notice their linguistic shortcomings. In contrast, L2 speakers at the higher end of the spectrum, who are conservative in their self-assessment, might preoccupy themselves with linguistic issues which are fairly inconsequential to their performance. In the end, inaccurate self-assessment might lead most L2 speakers, regardless of their ability level, to engage in skill-inappropriate linguistic choices, such as attending to some aspects of language (e.g., segmentals) at the expense of others (e.g., suprasegmentals), or in real-world behaviors, such as engaging in learning experiences which are inappropriate for their level. Clearly, the relationship between L2 users' self-assessment of their linguistic abilities and the extent to which they benefit from interaction needs to be investigated further.

Taken together, the current findings point to the inherent complexity of the relationship between language users' perceived and actual L2 performance. However, at least some of the variability in self-assessment could be due to measurement error, such as the tendency for scores to regress to the mean (Kruger & Mueller, 2002) or to test-takers' (or indeed teacher raters') bias to respond to traits or phenomena other than those targeted, to portray themselves in a positive light, or to be influenced by task difficulty (Bachman & Palmer, 1989; Heilenmann, 1990). Additionally, some of the variability in self-assessment may have also stemmed from methodological differences in how speakers and listeners assessed speech in this study. While the listeners were given multiple opportunities to listen to audio recordings to assess the speakers' accent and comprehensibility, the speakers had no access to their own performance or to the performance by their peers. Put differently, whereas the listeners could engage in norm-referenced assessment (i.e., compare one speaker to another speaker, with an order effect controlled through randomization), the speakers, who completed self-assessments without access to peer performances, could not evaluate their own performance relative to those of other L2 users. L2 speakers thus could not make immediate use of a reference sample of other speakers to mediate their use of the scale. Therefore, future research could further examine the relationship between L2 users' self- and peer-assessments, making their assessment procedure more similar to that of the listeners.

These methodological issues notwithstanding, consistently documented failures for people to accurately assess their performance, traceable to the lack of a threshold level of ability in a given skill, suggest that self-assessment is driven by entrenched, preconceived self-views (Carter & Dunning, 2008; Ehrlinger & Dunning, 2003). Critcher and Dunning (2009) proposed that such self-views influence people's performance evaluations by biasing them to experience tasks in a particular (excessively favorable) way. In their study, college students performing several tests were asked not only to estimate their test performance but also, when completing each test item, to self-rate how long they worked on it, how difficult it was, or how much they guessed. These researchers showed that students' self-assessment was unrelated to their actual performance but instead was associated with their perceived task experiences. Those who overestimated their performance, compared to those who did not, perceived their time on task as being shorter and the task itself as less effortful, and felt more familiar with test content. Thus, inaccurate self-assessment, guided by top-down self-views, was mediated through specific bottom-up task experiences.

Although no link between self-assessment and task difficulty was found here, Critcher and Dunning's (2009) finding implies that entrenched, preconceived self-views responsible for inaccurate self-assessment are ultimately traceable to peoples' real-world experiences with a given skill set. For a skill as ill-defined and complex as L2 pronunciation, where speakers might succeed in communication despite a noticeable accent or through the use of such strategies as gesturing, avoidance, or circumlocution to convey a message, and without interlocutors' feedback focusing specifically on speech perception and production, speakers might develop an overly positive view of their speaking experiences, thus reinforcing inaccurate self-views responsible for inaccurate self-assessment. Accurate self-assessment, then, may be tied to clear, unambiguous learning experiences, an issue discussed further as part of the implications.

Linguistic basis of self-assessment

Besides documenting inaccuracies in L2 speakers' self-assessment of their L2 accent and comprehensibility, one novel finding of this study was that such inaccuracies were only tied to segmental, suprasegmental, and fluency aspects of L2 speech (see Table 4). This result stands in stark contrast to research showing that, for native-speaking listeners, accent and comprehensibility are associated with different linguistic dimensions of speech (Crowther et al., 2014; Derwing et al., 2004; Tajima, Port, & Dalby, 1997; Trofimovich & Isaacs, 2012). L2 accent is predominantly tied to segmental and suprasegmental aspects of L2 speech as well as fluency while L2 comprehensibility, in addition to these factors, is linked to several other domains, including lexicon, grammar, and discourse structure. Thus, while L2 speakers' self-ratings of accent were inaccurate, compared to native listener ratings, they were based on the same linguistic dimensions that are used by native-speaking listeners to judge L2 accent (e.g., Crowther et al., 2014; Trofimovich & Isaacs, 2012). However, L2 speakers' inaccurate self-ratings of comprehensibility were linked only to segmental, suprasegmental, and fluency variables, whereas native-speaking listeners consider a wider range of linguistic factors in judging comprehensibility. In essence, L2 speakers appear to be unaware which linguistic factors, besides a few segmental and suprasegmental issues, make L2 speech comprehensible for the listener. This finding is striking given that comprehensible speech, rather than accent reduction, is considered to be more useful as a learning and teaching goal (Derwing & Munro, 2009; Levis, 2005). As was argued previously, this lack of awareness, compounded by the complexity of L2 pronunciation (and particularly comprehensibility) as a skill, would only exaggerate L2 speakers' inaccurate self-assessment.

Although investigating L1 effects on speakers' self-assessment was not the primary goal of this study, examining four L1-based groups separately in Experiment 2 proved advantageous, since nearly all theoretical frameworks of L2 speech learning predict L1-specific influences on L2 production (e.g., Eckman, 2004; Flege, 2003). Though sample sizes of individual groups were small, there were several L1-specific relationships between speakers' self-assessment of accent and comprehensibility and linguistic characteristics of their speech. For example, while inaccurate self-assessment of accent was linked to segmental accuracy for most groups, consistent with prior research (Dlaska & Krekeler, 2008), segmental issues had the strongest association for the Chinese speakers. This likely stems from the challenge that segmental production in English poses to these speakers, with many substitutions and errors of syllable structure in their speech (Anderson-Hsieh et al., 1992). For Hindi/Urdu speakers, segmentals, word stress, intonation, and rhythm were all strongly associated with their inaccurate self-assessment of comprehensibility. At least some of these associations might be specific to how intonation and pitch are used in Hindi/Urdu as compared to how they are used in English (Shackle, 2001). What is crucial is that these potentially L1-specific aspects of speech linked to L2 speakers' inaccurate self-assessment may not necessarily be those that actually matter for listener perception of accent and comprehensibility. In fact, in a companion study (Crowther et al., 2014), Hindi/Urdu speakers' comprehensibility for native-speaking listeners was associated with lexical, grammatical, and discourse-based aspects of their speech, rather than with any segmental, suprasegmental, or fluency variables. Put differently, the linguistic characteristics which are linked to L2 speakers' inaccurate self-assessment may not be the same characteristics which are actually important for native-speaking listeners' assessment. Although thought-provoking, these findings must be confirmed in future research, with the goal of disentangling possible culture- from language-specific effects on L2 speakers' inaccurate self-assessment, especially because speakers' culture and linguistic background were confounded in this study.

Implications and Future Research

One important implication of the current findings pertains to the issue of helping L2 speakers align or calibrate their self-assessment with their actual performance. This issue is crucial if indeed accurate self-assessment underlies L2 development, either through its impact on what L2 speakers attend to in the input they receive, or through its influence on real-word decision-making, for example, with L2 speakers engaging in learning experiences which are skill-inappropriate or abstaining from experiences which are beneficial. It appears that accurate self-assessment is not a simple matter of accruing skill-relevant experience and receiving feedback. For instance, students' semester-long experience with a subject in an undergraduate educational psychology course, given repeated testing and self-assessment opportunities, had little positive impact on their self-assessment accuracy, such that poor performers remained highly overconfident (Hacker, Bol, Horgan, & Rakow, 2000). Similarly, Simons (2013) showed that professional bridge players remained overconfident in their predictions of game success even though they consistently received feedback about their performance. In an extensive literature review, Dunning et al. (2004) offered several possible methods of improving self-assessment skills. Such methods include the use of self-testing exercises interspersed throughout the course, reviews of past performance, benchmarking (students creating an agreed-upon set of standards), and peer assessment. With respect to self-assessment of L2 speech, the effectiveness of these methods remains to be investigated in future research.

Last but not least, besides cognitive (e.g., Critcher & Dunning, 2009), motivational (e.g., Guenther & Alicke, 2010), or instructional (e.g., Dunning et al., 2004) variables contributing to inaccurate self-assessment, it might be important to consider social-psychological factors linked to overconfidence. For example, given the gender imbalance in our sample, it would be important to determine how speakers' gender impacts their self-views, as males and females tend to differ in their self-assessment behaviours (Dunning et al., 2003). Similarly, Fay, Jordan, and Ehrlinger (2012) recently reviewed evidence suggesting that socially-construed norms might encourage overconfident self-assessment behaviors. These researchers argued that social norms of preferring positive over negative feedback (e.g., Brown & Levinson, 1987), which may be culturally-mediated (e.g., Heine, Kitayama, & Lehman, 2001), and an emphasis on positive emotions such as joy, love, and pride over negative emotions (e.g., Eid & Diener, 2001) might actually rob people of the crucial information they require to create accurate self-awareness. Such social and emotional influences on self-assessment, particularly for L2 speech development in various contexts (e.g., classroom, study abroad) and for various types of learners (e.g., children, adults), are interesting areas for future research into the relationship between self- and other-assessment of L2 speech.

Notes

1. As some of the most commonly cited constructs in L2 pronunciation research (see Derwing & Munro, 2009), accentedness and comprehensibility are also associated with competing learning and teaching goals, namely, a focus on accent reduction versus understanding (Levis, 2005), and are frequently referred to in L2 oral proficiency scales (Isaacs & Trofimovich, 2012).

2. 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 with various linguistic dimensions of L2 speech. The Kaiser-Meyer-Oklin test and Bartlett's test of sphericity are used to test the assumption of factorability for a principal component analysis. These tests ensure that an appropriate level of correlations exists between variables to effectively run such an analysis.

3. Similar regression analyses carried out to examine how the PCA phonology and lexicogrammar factors predict accent and comprehensibility ratings, as judged by native-speaking listeners, showed that only the phonology factor was a significant predictor of accent, $R^2 = .74$, $B = 1.24$, *95% CI* = [1.04, 1.43], $t = 12.66$, $p < .0001$. In contrast, both the phonology factor, $R^2 = .56$, $B = .80$, *95% CI* = [.59, 1.00], $t = 7.71$, $p < .0001$, and the lexicogrammar factor, $R^2 = .11$, $B = .44$, *95% CI* = [.23, .65], $t = 4.25$, $p < .0001$, emerged as significant predictors of comprehensibility. Because these listener-based findings differed from the results of regression analyses targeting speakers' overconfidence scores, which incorporate self-ratings, it appears that the two linguistic dimensions of speech (phonology, lexicogrammar) contributed differently to self- versus listener-rated accent and comprehensibility.

4. One-way ANOVAs carried out to compare listeners' accent and comprehensibility scores for each speaker group supported this interpretation. There was a significant *F*-ratio for both accent, $F(3, 52) = 6.21$, $p = .001$, $\eta_p^2 = .26$, and comprehensibility, $F(3, 52) = 6.11$, $p = .001$, $\eta_p^2 = .26$, and Bonferroni-corrected post-hoc tests further showed that the Chinese group

was significantly more accented than the Farsi (*p* = .03) and Romance (*p* = .001) groups, and was significantly less comprehensible than the other three groups (*p* < .03).

5. An anonymous reviewer raised the intriguing possibility that the relationship between overconfidence in accent and comprehensibility ratings and linguistic dimensions of L2 speech could be related to L2 speakers showing different degrees of sensitivity to the phonology and lexicogrammar factors at different levels of L2 ability. In essence, less skilled L2 speakers (for whom phonology categories might pose a problem) may engage in overconfident rating behaviours based on other dimensions of their speech (such as lexicogrammar), while more skilled L2 speakers (for whom phonology might not pose considerable problems) may consider both phonology and lexicogrammar aspects of their speech for self-rating, leading to underconfident rating behaviours. As suggested by this reviewer, these interesting possibilities need to be explored in future research, for example, by having L2 speakers themselves evaluate their speech for multiple linguistic categories, with the goal of identifying which linguistic dimensions of speech L2 speakers heed at different levels of L2 ability.

References

Alderson, C. J. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. London: Continuum.

Ammar, A., & Spada, N. (2006). One size fits all? Recasts, prompts, and L2 learning. *Studies in Second Language Acquisition*, *28*, 543-574. doi:10.1017/S0272263106060268.

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. doi:10.1111/j.1467-1770.1992.tb01043.x

Bachman, L. F., & Palmer, A. S. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, *6*, 14-29. doi:10.1177/026553228900600104

Blanche, P. & Merino, B. J. (1989). Self assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, *39*, 313-340. doi:10.1111/j.1467-1770.1989.tb00595.x

Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System*, *40*, 144-160. doi:10.1016/j.system.2012.01.003

Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage.* Cambridge, UK: Cambridge University Press.

Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*, 60-77. doi:10.1037/0022-3514.90.1.60

Carter, T. J., & Dunning, D. (2008). Faulty self-assessment: Why evaluating one's own competence is an intrinsically difficult task. *Social and Personality Psychology Compass*, *2*, 346-360. doi:10.1111/j.1751-9004.2007.00031.x

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. doi:10.1016/S0346-251X(00)00036-1

Critcher, C. R., & Dunning, D. (2009). How chronic self-views influence (and mislead) self-assessments of task performance: Self-views shape bottom-up experiences with the task. *Journal of Personality and Social Psychology*, *97*, 931-945. doi:10.1037/a0017452

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2014). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*. doi:10.1093/applin/amt056

Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2014). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*. Published online 28 October 2014. doi:10.1002/tesq.203

Davidson, F., & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing*, *2*, 164-179. doi:10.1177/026553228500200205

de Saint Léger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals*, *42*, 158-178. doi:10.1111/j.1944-9720.2009.01013.x

Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*, 476-490. doi:10.1017 /S026144480800551X

Derwing, T. M., & Rossiter, M. J. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System*, *30*, 155-166. doi:10.1016/S0346-251X(02)00012-X

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. doi:10.1111/j.1467-9922.2004.00282.x

Dlaska, A., & Krekeler, C. (2008). Self-assessment of pronunciation. *System*, *36*, 506-516. doi:10.1016/j.system.2008.03.003

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself.* New York: Psychology Press.

Dunning, D., Heath, C., & Suls, J. M. (2004). Flawed self-assessment: Implications for health, education, and the workplace. *Psychological Science in the Public Interest*, *5*, 69-106. doi:10.1111/j.1529-1006.2004.00018.x

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*, 83-87. doi:10.1111/1467-8721.01235

Dunning, D., Meyerowitz, J. A., & Holzberg, A. D. (1989). Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of Personality and Social Psychology*, *57*, 1082-1090. doi:10.1037/0022-3514.57.6.1082

Eckman, F. (2004). From phonemic differences to constraint rankings: Research on second language phonology. *Studies in Second Language Acquisition*, *26*, 513-549. doi:10.1017/S027226310404001X

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology*, *84*, 5-17. doi:10.1037/0022-3514.84.1.5

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further exploration of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*, 98-121. doi:10.1016/j.obhdp.2007.05.002

Eid, M. & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology*, *81*, 869-885. doi:10.1037/0022-3514.81.5.869

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Education Research*, *59*, 395-430. doi:10.3102/00346543059004395

Fay, A. J., Jordan, A. H., & Ehrlinger, J. (2012). How social norms promote misleading social feedback and inaccurate self-assessment. *Social and Personality Psychology Compass*, *6*, 206-216. doi:10.1111/j.1751-9004.2011.00420.x

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Thousand Oaks, CA: Sage.

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. doi:10.1016/0167-6393(94)00044-B

Foote, J. A., Trofimovich, P., Collins, L., & Soler Urzúa, F. (2013). Pronunciation teaching practices in communicative ESL classes. *The Language Learning Journal*. Published online 16 April 2013. doi:10.1080/09571736.2013.784345

Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology, 62*, 378-391. doi:10.1037/0022-3514.62.3.378

Glover, P. (2011). Using CEFR level descriptors to raise university students' awareness of their speaking skills. *Language Awareness*, *20*, 121-133. doi:10.1080/09658416.2011.555556

Guenther, C. L., & Alicke, M. D. (2010). Deconstructing the better-than-average effect. *Journal of Personality and Social Psychology*, *99*, 755-770. doi:10.1037/a0020959

Hacker, D. J., Bol, L., Horgan, D., & Rakow, E. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*, 160-170. doi:10.1037/0022-0663.92.1.160

Harris, M. M., & Schaubroek, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology*, *41*, 43-62. doi:10.1111/j.1744-6570.1998.tb00631.x

Hayes, A. F., & Dunning, D. (1997). Construal processes and trait ambiguity: Implications for self-peer agreement in personality judgment. *Journal of Personality and Social Psychology*, *72,* 664-677. doi:10.1037/0022-3514.72.3.664

Heilenman, L. K. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing*, *7*, 174-201. doi:10.1177/026553229000700004

Heine, S. J., Kitayama, S., & Lehman, D. R. (2001). Cultural differences in self-evaluation: Japanese readily accept negative self-relevant information. *Journal of Cross-Cultural Psychology*, *32*, 434-443. doi:10.1177/0022022101032004004

Hodges, B., Regehr, G., & Martin, D. (2001). Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine, 76*, S87-S89.

Hutcheson, G. D., & Sofroniou, N. (1999). *The multivariate social scientist*. London: Sage.

Isaacs, T., & Trofimovich, P. (2011). *International students at Canadian universities: Validating a pedagogically-oriented pronunciation scale*. Unpublished corpus of second language speech.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility: Identifying the linguistic influences on listeners' L2 comprehensibility ratings. *Studies in Second Language Acquisition*, *34*, 475-505. doi:10.1017/S0272263112000150

Janssen van Dieten. A.-M. (1989). The development of a test of Dutch as a second language: The validity of self-assessment by inexperienced subjects. *Language Testing*, *6*, 30-46. doi:10.1177/026553228900600105

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.

King, C. (1994). *One language, two scripts: The Hindi movement in nineteenth century North India*. Delhi: Oxford University Press.

Klin, C. M., Guzman, A. E., & Levine, W. H. (1997). Knowing that you don't know: Metamemory and discourse processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1378-1393. doi:10.1037/0278-7393.23.6.1378

Kruger, J., & Dunning, D. (1999). Unskilled or unaware of it: Difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121-1134. doi:10.1037/0022-3514.77.6.1121

Kruger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*, 180-188. doi:10.1037/0022-3514.82.2.180

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. doi:10.2307/3588485

Lichtenstein, S., & Fischoff, B. (1977). Do those who know more also know more about how much they know? *Organization Behavior and Human Performance*, *20*, 159-183. doi:10.1016/0030-5073(77)90001-0

Little, D. (2005). The Common European Framework and the European Language Portfolio: Involving learners and their judgements in the assessment process. *Language Testing*, *22*, 321-336. doi:10.1191/0265532205lt311oa

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413-468). New York: Academic Press.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, K., D. Coste, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39-52). Amsterdam: John Benjamins.

Lyster, R. (2001). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, *51*, 265-301. doi:10.1111/j.1467-1770.2001.tb00019.x

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, *67*, 280-296. doi:10.1037/0021-9010.67.3.280

Mackey, A. & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 407-452). Oxford: Oxford University Press.

Mattern, K. D., Burrus, J., & Shaw, E. (2010). When both skilled and unskilled are unaware: Consequences for academic performance. *Self and Identity*, *9*, 129-141. doi:10.1080/15298860802618963

Moyer, A. (2013). *Foreign accent: The phenomenon of non-native speech*. Cambridge: Cambridge University Press.

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

of foreign-accented English. *Journal of Phonetics*, *25*, 1-24. doi:10.1006/jpho.1996.0031

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*, 1-20. doi:10.1177/026553229801500101

Saito, K. (2013). Communicative focus on second language phonetic form: Teaching Japanese learners to perceive and produce English /ɹ/ without explicit instruction. *Applied Psycholinguistics*. Published online 28 April 2013. doi:10.1017/S0142716413000271

Saito, K., Trofimovich, P., & Isaacs, T. (forthcoming). Using listener judgements to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Segalowitz, N. (2010). *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* (pp. 310-324). Cambridge: Cambridge University Press.

Simons, D. J. (2013). Unskilled and optimistic: Overconfident predictions despite calibrated knowledge of relative skill. *Psychonomic Bulletin & Review*, *20*, 601-607.

Sinkavich, F. J. (1995). Performance and metamemory: Do students know what they don't know? *Instructional Psychology*, *22*, 77-87.

Smiljanic, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Linguistics and Language Compass*, *3*, 236-264. doi:10.1111/j.1749-818X.2008.00112.x

Stemler, S.E., & Tsai, J. (2008). Best practices in estimating interrater reliability. In J. Osborne (Ed.). *Best practices in quantitative meth*ods (pp. 29-49). Thousand Oaks, CA: Sage.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility

Tesser, A., & Rosen, S. (1975). The reluctance to transmit bad news. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 193-232). New York: Academic Press.

Tokumoto, M., & Shibata, M. (2011). Asian varieties of English: Attitudes towards pronunciation. *World Englishes*, *30*, 392-408. doi:10.1111/j.1467-971X.2011.01710.x

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition*, *15*, 905-916. doi:10.1017/S1366728912000168

Upshur, J. (1975). Objective evaluation of oral proficiency in the ESOL classroom. In L. Palmer & B. Spolsky (Eds.), *Papers on language testing* (pp. 1967-1974). Washington: TESOL.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*, 335-353. doi:10.1177/0265532210364406

Yule, G., Damico, J., & Hoffman, P. (1987). Learners in transition: Evidence from the interaction of accuracy and self-monitoring skill in a listening task. *Language Learning*, *37*, 511-521. doi:10.1111/j.1467-1770.1987.tb00582.x

Yule, G., Hoffman, P., & Damico, J. (1987). Paying attention to pronunciation: The role of self-monitoring in perception. *TESOL Quarterly, 21*, 765-768. doi:10.2307/3586994

Appendix

Training Materials and Onscreen Labels

A. Speech rating	
Vowel and consonant errors	This measure applies to individual sounds and refers to errors in the pronunciation of individual sounds within a word. These errors may affect both consonants and vowels: o Speaker says that but you hear *dat* o Speaker says pen but you hear *pin* Such errors also include the removal and additions of sounds: o Speaker says house but you hear *ouse* o Speaker says spray but you hear *supray*
1	⬡--⬡ 1000
frequent	infrequent or absent
Word stress errors	This measure applies to individual words that are longer than one syllable and refers to errors in the placement of stress in words with more than one syllable. These errors include misplaced stress: o *comPUter* is pronounced as *compuTER* o *FUture* is pronounced as *fuTURE* These errors also include absent stress, so all syllables sound the same: o *comPUter* is pronounced as *computer* o *FUture* is pronounced as *future*
1	⬡--⬡ 1000
frequent	infrequent or absent
Intonation	This measure applies to utterances longer than a single word and can be described as the melody of speech. It refers to natural movements of pitch as we produce utterances. o Pitch goes up in *Will you be home tomorrow↑?* o Pitch goes down in *Yeah, I'll stay at home↓* o Pitch goes down and up in *I'll stay at home↓↑... but only until 3*. Intonation should come across as natural and unforced.
1	⬡--⬡ 1000
unnatural	natural

Rhythm

This measure applies to utterances and refers to differences in stress (emphasis) between content and function (grammatical) words.

- In *they RAN to the STORE*, the words *ran* and *store* are all content words and therefore are stressed more than the words *they*, *to* and *the*, which are grammatical words.

Rhythm should sound and feel natural in speech.

1 ○---○ 1000

unnatural natural

Speech rate

This measure applies to utterances and describes how slowly or quickly someone speaks.

- Speaker can speak slowly with many pauses and hesitations.
- Speaker can speak very fast.
- Speakers can speak at a natural rate and can be comfortable to listen to.

1 ○---○ 1000

too slow or too fast optimal

B. Transcript rating

Lexical appropriateness

This measure applies to individual words and refers to a speaker's choice of words to accomplish a speaking task. Poor lexical choices include incorrect, inappropriate, and non-English words.

- *I drank coffee with my friends in a fancy French cafeteria.*
- *A man and a woman bumped into each other on a walkside.*

1 ○---○ 1000

many inappropriate words used consistently uses appropriate vocabulary

Lexical richness

This measure applies to individual words and refers to the sophistication of the vocabulary used by a speaker to discuss a particular topic. Lexical richness is poor if a speaker uses very simple words with little variety.

- More rich utterance: *The girl arrived home to find her dog overjoyed at her return she quickly realized that he was more likely excited for the cookie he was about to receive*.
- Less rich utterance: *The girl arrived home her dog was happy she arrived home and the dog was happy too because he could eat a cookie.*

1 ○---○ 1000

few, simple words used varied vocabulary

Grammatical accuracy

This measure applies to both individual words and utterances longer than a single word and refers to the number of grammar errors made by the speaker. These may include:

- Errors of word order: *What you are doing?*
- Errors in grammar endings: *She go to school every day.*
- Agreement errors: *I will stay there for five day*.

1 ○---○ 1000

poor grammar accuracy excellent grammar accuracy

Grammatical complexity

This measure applies to utterances that are longer than a single word and describes the complexity and sophistication of a speaker's grammar. Grammar is sophisticated if a speaker uses complex and elaborate structures and embeds shorter utterances within longer

	utterances. o More complex utterance: *The man that was wearing a black hat was greatly enjoying his coffee*. o Less complex utterance: *The man wore a black hat… and he enjoyed his coffee*.
	1 ⬡--⬡ 1000
simple grammar	elaborate grammar
Discourse richness	This measure applies to the entire narrative and describes how rich and sophisticated a speaker's narrative is. o Discourse richness is low if the narrative is simple, unnuanced, bare, and lacks sophisticated ideas or details. o Discourse richness is high if a speaker produces several distinct ideas or details in his or her narrative, so that the story sounds developed and sophisticated.
	1 ⬡--⬡ 1000
simple structure, few details	detailed and sophisticated

Table 1

Background Characteristics (Means and Standard Deviations) for Speakers in Experiment 2

Variable	Farsi	Chinese	Hindi/Urdu	Romance
Gender (m/f)	9/6	5/9	13/1	9/4
Age of arrival in Canada	25.2 (2.4)	20.9 (6.3)	23.1 (2.1)	21.0 (3.2)
Years of English study	8.5 (4.8)	10.2 (3.0)	13.6 (6.1)	11.2 (4.5)
TOEFL iBT total score	87.8 (7.1)	85.0 (6.4)	88.5 (10.9)	86.4 (12.9)
IELTS total score	6.8 (.4)	6.3 (.5)	6.7 (.7)	7.0 (.7)
English use outside school[a]	21.0 (34.1)	17.9 (17.2)	36.4 (23.4)	30.0 (33.9)

Note. [a]Self-rating on a 0-100% scale.

Table 2

Summary of a Two-Factor Solution Based on a Principal Component Analysis of the 10 Rated Linguistic Variables

Factor 1 (Phonology)	Word stress errors (.99), Intonation (.95), Segmental errors (.92), Rhythm (.89), Speech rate (.49)
Factor 2 (Lexicogrammar)	Discourse richness (.98), Grammatical complexity (.97), Lexical richness (.97), Grammatical accuracy (.79), Lexical appropriateness (.77), Speech rate (.57)

Note. All eigenvalues > 1.

Table 3

Results of Multiple Regression Analyses Using the Factors of Phonology and Lexicogrammar as Predictors of Overconfidence Scores in Accent and Comprehensibility

Criterion variable	Predictors	R^2	*B*	*95% CI*	*t*	*p*
Overconfidence (accent)	Phonology	.28	–.12	[–.18, –.07]	4.63	.001
Overconfidence (comprehensibility)	Phonology	.27	–.11	[–.17, –.06]	4.62	.001

Note. The variables entered into the regression equation were the two factors obtained in the PCA reported in Table 2. However, because only the phonology factor emerged as a significant predictor, only this factor is listed in the table.

Table 4

Pearson Correlations Between the Five Phonology Categories and Overconfidence Scores for Accent and Comprehensibility (n = 56)

Category	Overconfidence in accent	Overconfidence in comprehensibility
Segmentals	–.59***	–.54***
Word stress	–.48***	–.45***
Intonation	–.48***	–.51***
Rhythm	–.43***	–.52***
Speech rate	–.32*	–.40**

Note. *p < .05, **p < .01, ***p < .001.

Table 5

Descriptive Statistics for Overconfidence Scores in Accent and Comprehensibility by L1 Group

Group	Overconfidence (accent)					Overconfidence (comprehensibility)				
	M	*SD*	*95% CI*	*Min*	*Max*	*M*	*SD*	*95% CI*	*Min*	*Max*
Farsi	–.04	.24	[–.17, .10]	–.48	.22	–.14	.21	[–.26, –.02]	–.44	.26
Chinese	.12	.19	[.01, .23]	–.33	.41	.09	.27	[–.07, .24]	–.33	.44
Hindi/Urdu	.15	.23	[.02, .29]	–.26	.52	.02	.16	[–.08, .11]	–.26	.22
Romance	–.11	.16	[–.20, –.01]	–.37	.15	–.14	.13	[–.21, –.06]	–.41	.07

Table 6

Pearson Correlations Between the 10 Rated Phonology Categories and Overconfidence Scores for Accent and Comprehensibility by L1 Group

Category	Overconfidence (accent)				Overconfidence (comprehensibility)			
	Farsi	Chinese	Hindi	Romance	Farsi	Chinese	Hindi	Romance
Segmentals	–.59*	–.75**	–.58*	–.34	–.35	–.40	–.70**	–.50
Word stress	–.58*	–.52	–.39	–.37	–.31	–.42	–.62*	–.40
Intonation	–.49	–.62*	–.43	–.35	–.45	–.58*	–.72**	–.36
Rhythm	–.37	–.46	–.23	–.37	–.32	–.44	–.65**	–.39
Speech rate	–.10	–.29	–.41	–.35	–.24	–.31	–.35	–.37

Note. **p* < .05, ***p* < .01.

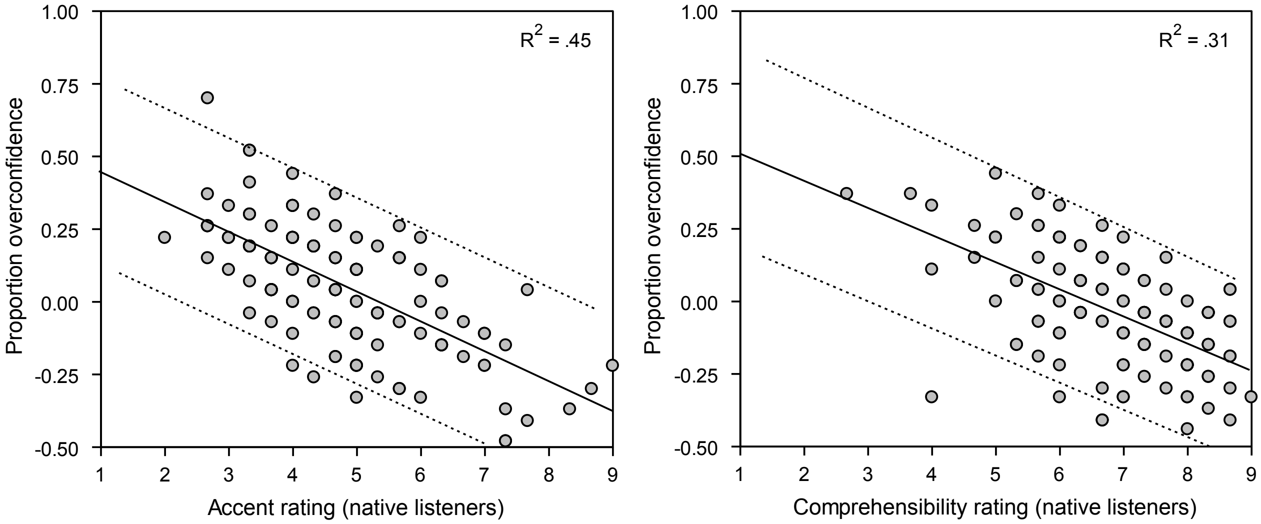

Figure 1. Associations between L2 speakers' (*n* = 134) overconfidence scores and their actual performance (as rated by native-speaking listeners) for accent (1 = *heavily accented*, 9 = *not accented at all*, left panel) and comprehensibility (1 = *hard to understand*, 9 = *easy to understand*, right panel), with regression lines showing the best fit to the data. Dotted lines designate 95% confidence intervals for linear regression.

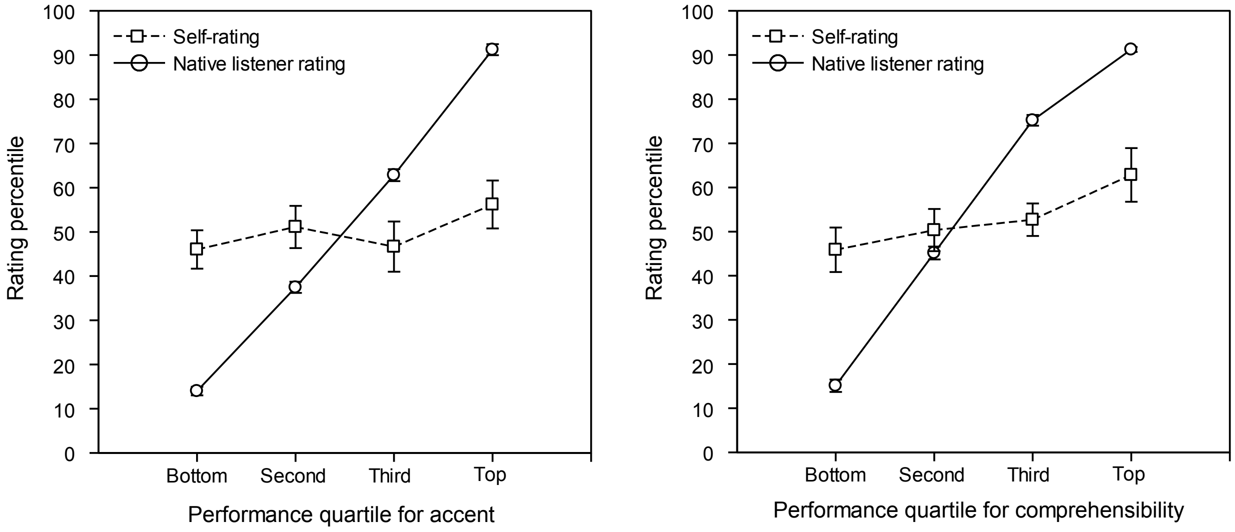

Figure 2. L2 speakers' (*n* = 134) percentile rankings for self- and listener-ratings of accent (left panel) and comprehensibility (right panel) as a function of listener-rated performance quartile (bottom to top 25%). Error bars enclose ±1SE.