



# On the Difficulty of Defining “Difficult” in Second-Language Vowel Acquisition

Murray J. Munro\*

Department of Linguistics, Simon Fraser University, Burnaby, BC, Canada

## OPEN ACCESS

### Edited by:

Andrew Sewell,  
Lingnan University, China

### Reviewed by:

Ocke-Schwen Bohn,  
Aarhus University, Denmark  
Charles Nagle,  
Iowa State University, United States

### \*Correspondence:

Murray J. Munro  
mjmunro@sfu.ca

### Specialty section:

This article was submitted to  
Language Sciences,  
a section of the journal  
Frontiers in Communication

**Received:** 08 December 2020

**Accepted:** 08 March 2021

**Published:** 05 August 2021

### Citation:

Munro MJ (2021) On the Difficulty of Defining “Difficult” in Second-Language Vowel Acquisition. *Front. Commun.* 6:639398. doi: 10.3389/fcomm.2021.639398

Hierarchies of difficulty in second-language (L2) phonology have long played a role in the postulation and evaluation of learning models. In L2 pronunciation teaching, hierarchies are assumed to be helpful in the development of instructional strategies based on anticipated areas of difficulty. This investigation addressed the practicality of defining a pedagogically useful hierarchy of difficulty for English tense and lax close vowels (/i ɪ u ʊ/) produced by Cantonese speakers. Unlike their English counterparts, Cantonese close tense-lax pairs are allophonic variants with [i u] occurring before alveolars and [ɪ ʊ] before velars. Each tense-lax pair represents a “phonemic split” in which members of a single L1 category are realized contrastively in L2. Despite evidence that English tense-lax distinctions are challenging for Cantonese speakers, no previous empirical work has closely considered the problem from the standpoint of vowel intelligibility across multiple phonetic contexts and in different words sharing the same rhyme. In a picture-based word-elicitation task, 18 Cantonese-speaking participants produced 31 high-frequency CV and CVC words. Vowels were evaluated for intelligibility by phonetically-trained judges. A series of mixed-effects binary logistic models were fitted to the scores, with vowel quality, phonetic context (rhyme) and word as factors, and length of Canadian residence and daily use of English as co-variables. As expected, the general hierarchy of difficulty for vowels that emerged (/i/ > /u/ > /ɪ/ > /ʊ/) was complicated by large differences across phonetic contexts. Results were not readily explicable in terms of transfer; moreover, different words with the same rhyme were not produced with equal intelligibility. The most serious modeling complication was the sizeable inter-speaker variability in difficulties, which could not be accounted for by model co-variables. Although some difficulties were roughly systematic at the group level, it is argued that establishing a pedagogically useful hierarchy on such data would prove intractable. Rather, L2 learners might be better served by assessment and instructional targeting of their individual problem areas than by a focus on errors predicted from hierarchies of difficulty.

**Keywords:** second language, intelligibility, vowels, Cantonese, phonology

## INTRODUCTION

A frequently cited goal of applied linguistics research is to unearth findings that advance classroom teaching practices. Although many research outcomes in the field are indeed pedagogically valuable, a connection between research and practice is not a *necessary* component of studies of second language (L2) learning. In fact, much L2 research is designed to resolve theoretical questions without direct relevance to instruction (White, 2018). Studies conducted without explicit pedagogical motivation may, of course, yield incidental implications for teaching, but that outcome is more likely the exception than the rule.

Specifically within the realm of pronunciation, Brinton (2018, p. 283) observes that “the nature of second language pronunciation research often precludes its application to the classroom.” In addition to the theoretical emphasis mentioned above, at least two further reasons may be offered for the lack of relevance she describes. One of these is a long-standing pre-occupation with native-like accuracy in assessments of L2 production, in accordance with the “nativeness principle” (Levis, 2005, 2020). In their historical overview, Munro and Derwing (2011) pointed to a dearth of empirical pronunciation studies motivated by the opposing “intelligibility principle,” despite repeated calls for an instructional focus on intelligible L2 speech going back at least as far as Sweet (1900). In a narrative summary of 75 mainly twenty-first century studies of instructional effectiveness, Thomson and Derwing (2015) reported that 63% were aligned with the nativeness principle. That orientation emphasizes *accentedness*, the degree to which speech is judged to differ from some particular variety (Munro and Derwing, 1995, 2020). For pedagogy, such a focus is deeply problematic in that empirical data conclusively demonstrate that native-sounding pronunciation is neither necessary for effective communication, nor likely to be achievable in the majority of L2 learners (for an overview, see Derwing and Munro, 2015; for a replication study, see Nagle and Huensch, 2020). Rather, even heavily-accented L2 speech can be highly *intelligible* (understood as the speaker intends) and highly *comprehensible* (easy to process). Given the incongruence between accentedness and the other dimensions, findings from nativeness-driven research are apt to create misapprehensions among teachers and pronunciation researchers. For instance, an intervention that leads to no change in learners’ accentedness might be wrongly interpreted as evidence of the ineffectiveness of instruction, even though empirical findings show that L2 speech can become more comprehensible through instruction despite no change in accentedness (e.g., Derwing et al., 1998; Derwing and Rossiter, 2003).

A partial explanation for the quasi-independence of accentedness from the other two dimensions comes from a functional load (FL) account of error gravity (Catford, 1987; Levis and Cortes, 2008; Sewell, 2017), which assumes that certain linguistic contrasts play a greater role in distinguishing meanings than do other contrasts. Although FL itself is not a focus of the present study, the vowel sounds at issue (English close vowels) provide a useful example of the concept. The English /i/-/ɪ/

contrast is considered to have a high FL for several reasons, including that it distinguishes a large number of minimal pairs, such as *seat-sit*, *feel-fill*, and *leave-live*; that many of these pairs are easily confusable because of identical syntactic functions (*leave* and *live* can both function as verbs); and that the words at issue are of relatively high frequency. In contrast, the /u/-/ʊ/ distinction has a low FL because of its relative rarity, especially in commonly-used, confusable words. Examples include *Luke-look* (unlikely to be confused), *wooed-wood* (unlikely to be confused; first item is rare), and *tuque-took* (unlikely to be confused; first item, a type of hat, is largely unknown outside Canada). Findings from Munro and Derwing (2006) suggest that a failure to produce high FL distinctions can reduce comprehensibility more than a failure to produce low FL distinctions, yet the two types appear to increase accentedness to a similar degree. It follows that if intelligibility is an instructional goal, prioritizing high FL difficulties is likely to be more effective than an approach that treats all difficulties as equally problematic.

Returning to Brinton’s (2018) point mentioned above, an additional factor in the lack of relevance of pronunciation research has been a preoccupation with differences in group-level performance in L2 speech production coupled with insufficient attention to inter-learner variability. In spite of twenty-first-century, rethinking of the uses of statistics in the social sciences (Larson-Hall, 2015), quantitative applied linguistics research relies heavily on null hypothesis statistical testing (NHST), which entails reporting of means, dispersion and effect sizes to evaluate between-group differences. In spite of the theoretical value of capturing such tendencies in pronunciation data, that emphasis appears to be at odds both with common anecdotal reports and with longitudinal data (Munro et al., 2015) indicating large inter-learner differences in pronunciation learning trajectories. Recent work by Wade et al. (2020) illustrates how an excessive focus on group means “masks the complexity of phonetic behaviors.” Their native English VOT data showed that individual departures from group performance were often stable, indicating that between-speaker variability was not simply “noise,” but reflected systematic individual differences from group norms that were not explicable on linguistic or sociolinguistic grounds. One might expect a parallel phenomenon in L2 production.

One issue in L2 pronunciation research that is commonly cited as pedagogically relevant is the determination of segmental “hierarchies of difficulty.” A long-standing assumption is that theory-driven research should facilitate advance prediction of the L2 segmental difficulties faced by learners from particular L1s, and that teachers would benefit from such knowledge. In his skeptical coverage, Walz (1980), for instance, commented on the view among applied linguists that such research would lay the groundwork for good teaching materials and techniques. And years earlier, Moulton (1962) had argued that theoretical explanations of the sources of L2 pronunciation errors would lead to improved instruction. Nearly half a century later, research on L2 segmental difficulties experienced by particular L1 groups is ubiquitous and has expanded beyond L2 English. Even when only vowels are considered, the diverse range of L2s covered includes Danish (Bohn and Garibaldi, 2017), Dutch (Burgos et al., 2014),

German (O'Brien and Smith, 2010; Darcy and Krüger, 2012; Nimz and Khattab, 2020), Polish (Sypiańska, 2016) and Japanese (Okuno and Hardison, 2016). Such studies attest to the health of the field of L2 pronunciation research and promise advancements in our understanding of the process of L2 speech acquisition. At the same time, judging by the continued popularity of texts presenting lists of difficult L2 phones organized by L1 (e.g., for English, Swan and Smith, 2001; Nilsen and Nilsen, 2010), there is still considerable enthusiasm among teachers and learners for predicting learner difficulties in advance (Munro, 2018). Whether that enthusiasm is warranted, however, is open to question.

Numerous accounts of the sources of difficulty in L2 phonological acquisition have been proposed. Among these, the best known concept is *transfer*, widely recognized as a fundamental influence on L2 speech production (Archibald, 2017). Attempts to characterize L2 phonology strictly in terms of transfer took the form of the strong version of the Contrastive Analysis Hypothesis (CAH). However, that approach failed at the levels of both segments (Brière, 1966; Wardhaugh, 1970) and prosody (Liu, 2021). In his detailed survey of twentieth-century L2 phonology, Eckman (2004) discusses how research has been extended to encompass other predictive factors as a way of improving upon transfer-based analyses. At the linguistic level these include similarity, markedness, and phonological constraints. It is also recognized that individual differences in learning are worthy of close attention (Edwards, 2017), and considerable work goes beyond linguistic considerations to include an array of psycho-social variables such as age of second language learning (AOL) and language experience (e.g., the Speech Learning Model; Flege, 1995), group affiliation (Gatbonton et al., 2005), cognitive processing ability (Darcy et al., 2015) and motivation (Saito et al., 2017; Nagle, 2018).

The present study concerns aspects of English close vowel acquisition that pose difficulty for Cantonese-speaking immigrants in Canada. In particular, I ask whether it is feasible to identify a pedagogically useful hierarchy of difficulty, given two contemporary concerns described above: the intelligibility principle and individual learner differences. For the purposes of this investigation it is irrelevant whether any particular theoretical model proves statistically more accurate than another. Rather, no specific theoretical framework will be assumed. Instead I use a *post-hoc* approach similar to that of Walz (1980) without seeking detailed explanations for patterns of performance. In this case, L2 productions of vowels in selected phonetic contexts are elicited from a cohort of Cantonese-speaking immigrants. These are evaluated to determine the relative difficulty of the vowels in the contexts at issue. Then I consider whether the *post-hoc* patterns of difficulty that emerge can be described in such a way that a language teacher could hypothetically apply the results directly to instruction given to Cantonese-speaking learners of English. In simple terms, my concern is with the “what,” rather than the “why” of L2 vowel production.

Findings from some previous work on L2 vowel intelligibility call into question whether the proposed hierarchy is achievable. For instance, Munro et al. (1996) examined English vowel production by 240 Italian speakers. Their focus was the impact on production of the speakers' AOL (ranging from early childhood to young adulthood) with respect to both vowel nativeness and vowel intelligibility. The results for nativeness were theoretically interesting in that a strong, linear AOL effect emerged, with little indication of native-like productions among late teen and young adult learners. However, the AOL effect was much weaker for intelligibility. Canadian English /oʊ/ and /ɒ/, for instance, were produced with high intelligibility by nearly all AOL groups, despite their being judged as heavily “accented” in many cases. Moreover, individual speakers' performance on vowels varied widely, leading the authors to conclude that strong generalizations about Italian speakers' areas of difficulty could not be made without ignoring important individual differences.

Additional concerns arise from a longitudinal investigation of English vowel intelligibility in Mandarin and Slavic Language speakers (Munro and Derwing, 2008). Even without focused pronunciation instruction, vowel intelligibility increased for both groups during the first 6 to 8 months of Canadian residence. However, improvement was not uniform either across vowels or within groups. Instead, learning trajectories varied from one learner to another, and no full explanation for the differences could be offered.

Another potential complexity in developing a difficulty hierarchy concerns the generalizability of vowel learning across different phonetic contexts and in similar phonetic contexts across different words. While the simplest possible hierarchy would be a list of vowels in order of difficulty, evidence suggests that this is not achievable. For instance, in Thomson's (2011) study of the benefits of perceptual training on vowel production, the effects of the intervention, assessed in terms of improved intelligibility, generalized to some, but not all phonetic contexts. Interestingly, the context at issue was the consonant *preceding* the trained vowel. That outcome supports the view that L2 vowel intelligibility is at least to some degree context-dependent, and that a difficulty hierarchy would have to specify different levels of difficulty according to context. A related issue is whether generalization can occur across the similar phonetic contexts of rhyming words, such as *seat*, *heat*, and *feet*. That question, which has yet to be closely examined in an intelligibility study, will be given close attention here.

The present study uses a descriptive approach to address the problem areas identified above: it assesses vowel intelligibility across a cohort of L2 English speakers from a shared L1 background (Cantonese), taking into account different phonetic contexts and different words that are phonetically similar. This close examination of individual learners' productions will provide insight into the feasibility of characterizing learners' performance in terms of hierarchy of vowel difficulty. Although intelligibility will be considered at the group level, considerable attention will be focused on how well group-based generalizations may obscure important individual variability in individual speakers. In particular, highly idiosyncratic

performance would pose a serious problem for establishment of a useful hierarchy.

## METHODS

### Focus of the Investigation

The vowels / i ɪ u ʊ / were selected for study because of their different distributions in English and Cantonese. While the four are phonemically contrastive on the basis of quality in Canadian English, Cantonese makes only a two-way phonemic distinction in which tense and lax vowels are allophonic variants. Chan and Li (2000) state that front /i/ is realized as [i:] before labials and alveolars and as [ɪ] before velars, and that back /u/ is produced as [u:] before alveolars and as [ʊ] before velars. Roughly, then, the English rhymes /it/, /ut/, /ik/, /ʊk/ match sequences that occur in Cantonese, while /ɪt/, /ʊt/, /ɪk/, /ʊk/ do not. From the standpoint of classical Contrastive Analysis, the problem faced by learners is referred to as a “split category” or “allophonic split” (Brière, 1966; Eckman et al., 2003). In addition, Cantonese syllable codas are restricted to nasals and voiceless plosives. Therefore, no English vowel + /d/ rhymes have matches in Cantonese. Finally, Zee (1991) observes that both /ɪ/ and /ʊ/ are relatively lowered, suggesting that they are close but not exact matches for their Canadian English counterparts. All three of these differences will be considered in this investigation.

Research on L2 English vowel production by Cantonese speakers points to difficulties with close vowels. Meng et al. (2007), for example, note that common errors include pronunciation of /ɪ/ as /i/ in words like *sit* and of /ʊ/ as /u/ in *full*. However, intervention studies indicate that adult L2 speakers from Cantonese backgrounds are capable of improved intelligibility after training. Wong (2015) reported more accurate production of close tense and lax vowels after high variability perceptual training (HVPT), though improvement was greater in the front pair than in the back one. Wong (2013) also found intervention to be beneficial, but neither study examined contextual effects closely, and neither considered the consistency of vowel accuracy across different words with identical rhymes (e.g., *seat*, *heat*).

### Speakers

English is well-established in Hong Kong, and the characteristics of the Hong Kong variety of English have been discussed by Hung (2000). Nonetheless, individual speakers' English proficiency varies widely. Some immigrants to Canada from Hong Kong have little experience using English for social or work-related communication and regard themselves as having insufficient English skills to function in Canadian society. They therefore view themselves as English learners and may choose to enroll in English classes after arrival. Speakers from that demographic were targeted for this investigation.

Speech data were collected from 18 native speakers of Cantonese (10 female, 8 male) recruited via post-secondary student e-mail lists and word-of-mouth on campuses in the lower mainland of British Columbia (i.e., Metro Vancouver).

Informational materials called for participants from Hong Kong who self-identified as second-language speakers of English, who were between 19 and 30 years of age with normal hearing, and who had arrived in Canada at age 15 or later. A CAD\$25 payment was given as an incentive. One participant who initially enrolled was dropped from the study (and replaced) for not meeting the age criterion.

Prior to recording, each speaker completed a language background questionnaire (LBQ). The speakers had been born and raised in Hong Kong and had moved to Canada at a mean age of arrival (AOA) of 18 years [range: 15–25 years]. Mean age at the time of the study was 23 years [range: 19–28], and length of residence (LOR) in Canada averaged 4.9 years [range: 9 months–6.9 years]. All had grown up in households with native Cantonese-speaking parents who used Cantonese as their household language. Though all speakers had studied English from the beginning of their schooling, as was typical in Hong Kong at the time, none reported regular use of English for social purposes until arrival in Canada. Eleven speakers had attended Canadian ESL classes, with a mean duration of 11.3 months, and two speakers reported completion of a 1–2 mo. course in English pronunciation. At the time of the study all had enough proficiency in English to be enrolled in English-speaking Canadian post-secondary institutions, where they were registered for degree credit. On the LBQ, each speaker completed a grid to estimate personal use of spoken English and Cantonese during each morning, afternoon and evening in a typical week. On average, the speakers reported using English 26% of the time [range: 0–93%]. The cases of 0% English use [ $n = 2$ ] are attributable to recruitment during the summer semester when some speakers were not enrolled in classes and reported exclusive use of their L1 at home and work.

For reliability assessment of the judges' evaluations (see below), recordings of two native English speakers (1 female, 1 male) were also used. These were randomly selected from a database of 18 speakers of Canadian English, also post-secondary students. All 20 speakers in the study passed a pure-tone hearing screen (250–4,000 Hz at 20 dB<sub>HL</sub>).

### Test Items

The test items were English CV(C) words covering the 14 actually-occurring English rhymes consisting of the vowels /i/, /ɪ/, /u/, and /ʊ/ in open syllables and in checked syllables before /t/, /k/, and /d/. These combinations were selected because they appear to be an optimal set for this study, given that Cantonese allows rough equivalents to some of them, but not to others. For instance, it is possible to examine whether the L2 speakers perform better on the “matching” /it/ rhyme than on “non-matching” /ɪt/. Target words with /d/ as a coda were also included. These are interesting in that final /d/ is absent in Cantonese, but rhymes containing (homorganic) final /t/ are possible. It is unknown whether the speakers might be able to generalize knowledge of Vt rhymes to rhymes matching in place of articulation but not in voicing. Although other syllable codas could have been included, the number of tokens needed to cover the above combinations is already large. Expanding the list would

**TABLE 1** | Target words from the picture-naming task.

Vowel	Coda	Words
/i/	#	key see tea
	/t/	feet, heat, seat
	/k/	cheek, speak
	/d/	feed, read
/ɪ/	/t/	hit, sit
	/k/	chick, kick, sick
	/d/	kid, lid
/u/	#	Sue, two
	/t/	boot, suit
	/k/	Luke, tuque
	/d/	food
/ʊ/	/t/	foot, put
	/k/	book, cook, look
	/d/	good, wood

add more work to an already onerous task for the judges (see section Judges' Evaluations).

To successfully elicit productions it was necessary to find familiar words that could be easily depicted in drawings. For that reason, it was not possible to generate a set of suitable words such that equal numbers of each rhyme were represented; nor was it possible to fully match words for initial consonants or to incorporate a complete set of minimal pairs. The final word list is given in **Table 1**.

## Speaking Tasks

During individual recording sessions, the speakers sat in an audiometric booth wearing a Shure Beta 54 head-mounted microphone connected to a Symetrix 302 microphone preamplifier and an HHB Professional digital recorder (CDR-830) set to a sampling rate of 44.1 kHz with 16-bit resolution.

A picture naming task was used to elicit the target words, so that the productions could be assumed to be based on stored representations developed through experience with English. It was expected that avoiding written forms would minimize orthographic influences. Participants named the target words from a randomly-ordered set of drawings, each accompanied by the first letter of the word as a cue. For instance, a drawing of a chair, along with "S" was used to elicit *seat*, and an arrow pointing to a foot with "F" indicated *foot*. The drawings were mounted on individual 8.5" by 11" cards that were randomized by shuffling. Each card included a stimulus number (1 to 31). During an unrecorded practice session, a research assistant went through the entire pack, and speakers guessed each target item. They then practiced producing the stimulus number and the target item in the following sentence frame: "Number \_\_. The next word is \_\_." The stimulus number was used for later sorting of the randomized recorded items. In case of a wrong guess, the speaker was required to guess again until the correct word was produced. *Sue* and *Luke* were depicted by a female and

male face, respectively, and the assistant explained that proper names beginning with "S" and "L" were required. During the practice elicitation it was discovered, as expected, that none of the Cantonese speakers were familiar with the word *tuque* (/tuk/, also spelled *toque*), even though this word is widely used in Canada to refer to a winter hat and was produced without hesitation by all native English speakers in the database. The research assistant therefore modeled the word and allowed each speaker to practice it a few times prior to recording.

On completion of the practice session, the pack of cards was shuffled. Each participant then recorded the full stimulus set three separate times, with a shuffling of the cards after each run-through. In case of a hesitation, a repetition was elicited (fewer than 1% of cases). During each run-through, extra items (not used in any analyses) were added to the beginning and end of the pack to minimize effects of list intonation in production. The picture elicitation task was followed by additional speaking and listening tasks to be reported elsewhere.

## Judges' Evaluations

The recorded words from all 20 speakers were extracted digitally and saved as peak-normalized, single-word audio files. Prior to evaluation by the judges, the author informally pre-screened the recordings and judged that some vowel productions resembled English vowels other than the actual target items. A majority of productions were heard to fall into one of the seven English categories / i ɪ eɪ ε u ʊ oʊ /. However, some tokens were ambiguous, falling between two of the categories, and a small number (<1%) were indeterminate.

The listener-judges were four linguistically-trained assistants, all familiar with IPA, and all having passed the same pure-tone hearing screen as the speakers. During individual listening sessions (nine per judge) held on different days, each judge heard the recorded words one at a time and identified each vowel in a forced-choice task (as in Munro and Derwing, 2008). The judges sat in a sound-treated room, each hearing a different random presentation via custom-designed playback software through AKG K141 professional studio headphones. They responded by clicking computer buttons marked with the phonetic symbols representing the seven vowel categories described above. An extra button marked "Other" was available for vowels not heard as one of the possible choices. Judges were advised that a vowel might sound foreign-accented, yet still clearly belong to one of the possible categories. In case of ambiguity, they were told to choose the vowel symbol that best represented the production. The task was self-paced: once a response was entered, the computer automatically played the next item. A practice session of about 5 min was given, during which the listening volume was adjusted to a comfortable level. Over the nine sessions, each judge provided a total of 1,860 responses.

## RESULTS

### Reliability of Judges

For assessment of inter- and intra-judge reliability and for the intelligibility analyses that follow, the judges' responses were coded as binary values to indicate on-target (i.e., "correct

**TABLE 2** | Type III tests of fixed effects (vowel).

Effect	df (Num)	df (Den)	F	Pr > F
Vowel	3	1,866	188.56	<0.0001
LOR	1	14.62	4.33	0.0554
% USE	1	14.84	2.79	0.1159

category”) or off-target (“incorrect category”) identifications. Inter-judge agreement was operationalized as the number of times the judges agreed with each other on whether or not a token was on-target. At least 3 out of 4 judges agreed on 91% of the Cantonese speakers’ tokens, with 4-way agreement on 70%. These results are very similar to those reported by Munro and Derwing (2008). Fleiss’ Kappa was computed at  $\kappa = 0.56$ ,  $p < 0.001$ , CI [0.542, 579], indicating moderate agreement according the Landis and Koch (1977) benchmarks. It should be noted that high reliability was not expected here because of the ambiguity in many of the productions that was noted above. That is, many productions seemed to straddle more than one vowel category such that a judge would have difficulty making a straightforward classification.

Intra-judge consistency was assessed by requiring the judges to re-evaluate 84 tokens during a separate listening session held on a different day after the original evaluations were completed. Consistency was computed by determining the percentage of times each judge evaluated items the same way (on-target/off-target) both times. Consistency ranged from 89 to 93% across the four judges. Given the ambiguous nature of some of the productions, this was deemed very acceptable.

## Intelligibility

For the Cantonese speakers’ vowels, the mean correct ID rate was 73%. For the native English tokens, which were included strictly as a means of verifying correct use of the vowel symbols, mean correct ID was 99%, suggesting a high level of accuracy in symbol use by the judges. In the analyses that follow, the results for only the Cantonese productions will be discussed.

On-target identifications on the open-syllable vowels /i/ and /u/ were 100%; therefore, scores on these items were excluded from statistical computations. The data from the checked-syllable productions were submitted to generalized mixed-effects modeling with Glimmix in SAS 9.4 (2018, SAS Institute, Cary NC), using the Kenward-Roger approximation for degrees of freedom. Speakers and judges were entered as random effects, with percentage of weekly English use as reported by participants in the LBQ (%USE) and length of residence in months (LOR) as co-variates. Vowel, Rhyme and Word were fixed effects. Because the fixed factors were not independent, it was necessary to compute a separate model for each one. Except where indicated otherwise, Tukey-Kramer *post-hoc* tests were used for follow-up analyses, with overall  $p < 0.05$  as the criterion for significance.

## Vowel Comparisons

For checked-syllable vowels, the order of intelligibility was /i/ (92%) > /u/ (82%) > /ʊ/ (59%) > /ɪ/ (51%). Mixed-effects

modeling (summarized in **Table 2**) yielded a significant effect of Vowel, with a non-significant contribution of %USE and a marginally non-significant contribution of LOR.

*Post hoc* tests revealed significant differences for all pairwise combinations, according to the ordering given above. In sum, the two lax vowels were produced significantly less intelligibly than their tense counterparts. For the tense pair, the front vowel was significantly more intelligible than the back vowel, while the reverse was true for the lax pair.

## Rhyme Comparisons

Performance on the checked-syllable rhymes is summarized in **Table 3**, according to speaker and rhyme. Speakers are ordered vertically from highest to lowest mean correct identifications, and rhymes are ordered from left to right. For individual speakers, intelligibility rates ranged from 89 to 58%, while the range for individual rhymes was from 97% for /id/ to only 33% for /ʊk/. Also indicated in the table (by “+”) are cases in which a particular speaker could be considered to have “acquired” a particular rhyme, according to the 80% correct criterion suggested by Carlisle (1998). Across speakers the total number of intelligible rhymes (second-to-last column) ranged from 10 to 3, out of a possible 12. The number of speakers reaching criterion on particular rhymes (second-to-last row) ranged from 17 for /id/ to just a single speaker for /ʊk/.

Model results for Rhyme, shown in **Table 4**, were parallel to those for the Vowel analysis. They indicated a significant effect of Rhyme, and non-significant contributions of LOR and %USE.

*Post hoc* comparisons indicated that performance on /id/ was significantly better than on all other rhymes except /ik/ while the mean score on /ʊk/ was significantly worse than on all other rhymes. The number of possible pairwise comparisons for the 12 rhymes is 66; however, only a subset of the outcomes—those most pertinent to the questions addressed in this study—will be considered here. First, **Table 5** compares performance on Vt and Vk rhymes, with the left column (matching VCs) showing rhymes that have approximate counterparts in Cantonese and the right column (non-matching) showing those that do not. In none of the 4 pairwise comparisons did performance on matching VCs statistically exceed scores on non-matching VCs. In fact, the significant differences that emerged in three comparisons all favored the non-matching VC, while for /ik/ and /t/, intelligibility was relatively low in both cases (52 and 44%), and the difference failed to reach significance.

**Table 6** shows performance on Vt and Vd rhymes for each of the four vowels. In three of the four cases, Vd performance significantly exceeded Vt, with no statistical difference for the fourth, /ut/ vs. /ud/.

## Word Comparisons

The results of the mixed-effects analysis for Word are given in **Table 7**. Again, these parallel the results of the previous modeling, with Word yielding a significant effect, and non-significant effects for the two co-variates.

The number of possible pairwise word comparisons (325) is very large, but the comparisons of interest for the purposes of this study are the smaller subset of words with identical

**TABLE 3 |** Intelligibility of vowels by rhyme and speaker.

SPKR	LOR <sup>a</sup>	%USE <sup>b</sup>	id	ik	uk	ʊd	it	ut	ʊt	ud	Id	Ik	It	ʊk	Acquired rhymes <sup>c</sup>	Mean %-ID
c009 <sup>ESL</sup>	6.8	72	+	+	+	+	+	+		+	+	+	+		10	89
c014	5.7	33	+	+	+		+	+	+	+	+			+	9	85
c018	6	16	+	+	+	+	+	+		+					7	84
c015	6.8	16	+	+	+	+	+		+		+		+		8	81
c016 <sup>ESL</sup>	3.4	32	+	+		+	+				+	+			6	79
c022 <sup>ESL</sup>	5.9	0	+	+	+	+	+				+		+		7	78
c012 <sup>ESL</sup>	6.8	93	+	+	+	+						+			5	78
c020 <sup>ESL</sup>	6.7	71	+	+	+	+	+		+						6	72
c019	3.1	48	+	+	+		+	+		+					6	72
c013 <sup>ESL</sup>	6.7	6		+		+			+						3	71
c021	6	2	+	+	+	+			+						5	69
c005 <sup>P</sup>	0.8	33	+	+	+	+	+	+		+					7	68
c004 <sup>ESL</sup>	3.8	0	+	+	+		+			+					5	68
c002 <sup>ESL</sup>	4.6	17	+	+	+	+			+						5	67
c003	2.7	6	+	+	+	+	+			+					6	65
c008 <sup>ESL</sup>	5.6	9	+	+	+	+		+							5	65
c007 <sup>ESL</sup>	4.9	4	+			+			+						3	64
c006 <sup>ESL,P</sup>	2.8	13	+		+					+					3	58
Total speakers			17	16	15	14	11	6	7	8	5	3	3	1		
Mean %-ID			97	93	88	86	81	70	69	62	55	52	44	33		

“+” indicates that the vowel in the rhyme was accurately identified at least 80% of the time by the four judges. <sup>a</sup>Length of residence in Canada in years. <sup>b</sup>Weekly use of English. <sup>c</sup>Total rhymes (maximum 12) showing 80% correct ID or higher. <sup>ESL</sup>Participant had received ESL instruction in Canada. <sup>P</sup>Participant had received focused instruction in English pronunciation.

**TABLE 4 |** Type III tests of fixed effects (rhyme).

Effect	df (Num)	df (Den)	F	Pr > F
Rhyme	11	1,858	80.18	<0.0001
LOR	1	14.66	4.36	0.0547
% USE	1	14.84	2.78	0.1167

**TABLE 5 |** Intelligibility of vowels in syllables according to approximate match or non-match of rhymes in English and Cantonese.

VC [matching]	Results*	VC [non-matching]
it (81)	<	ik (93)
Ik (52)	=	It (44)
ut (70)	<	uk (88)
ʊk (33)	<	ʊt (69)

\* < indicates that performance on the matching VC was significantly worse; = indicates no statistical difference between matching and non-matching.

rhymes. Accordingly, a total of 17 Bonferroni-adjusted *t*-tests (with overall *p* < 0.05) were performed, the results of which are summarized in **Table 8**. Here, the central issue is whether vowels in words with identical rhymes showed approximately equal levels of accuracy. As can be seen, nearly half (eight) of the pairwise comparisons yielded significant differences. The *heat* and *chick* vowels were produced less intelligibly than the vowels in the other two words with identical rhymes, and the *cheek*,

**TABLE 6 |** Intelligibility of vowels in Vt and Vd rhymes.

Vt	Results*	Vd
it (81)	<	id (97)
It (44)	<	Id (55)
ut (70)	=	ud (62)
ʊt (69)	<	ʊd (86)

\* < indicates that performance on the Vt target was significantly worse; = indicates no statistical difference between Vt and Vd targets.

**TABLE 7 |** Type III tests of fixed effects (word).

Effect	df (Num)	df (Den)	F	Pr > F
Word	25	1,844	37.08	<0.0001
LOR	1	14.67	4.37	0.0543
% USE	1	14.67	2.77	0.1171

*foot*, *Luke* and *lid* vowels were less intelligible than those in the rhyming words *speak*, *put*, *tuque*, and *kid*, respectively.

**Individual Speaker Performance**

Individual speaker data on words with identical rhymes revealed noteworthy differences in words across speakers. Illustrative examples are provided in **Figures 1–3**. **Figure 1** shows individual performance on *kid* and *lid*. Note that the group means

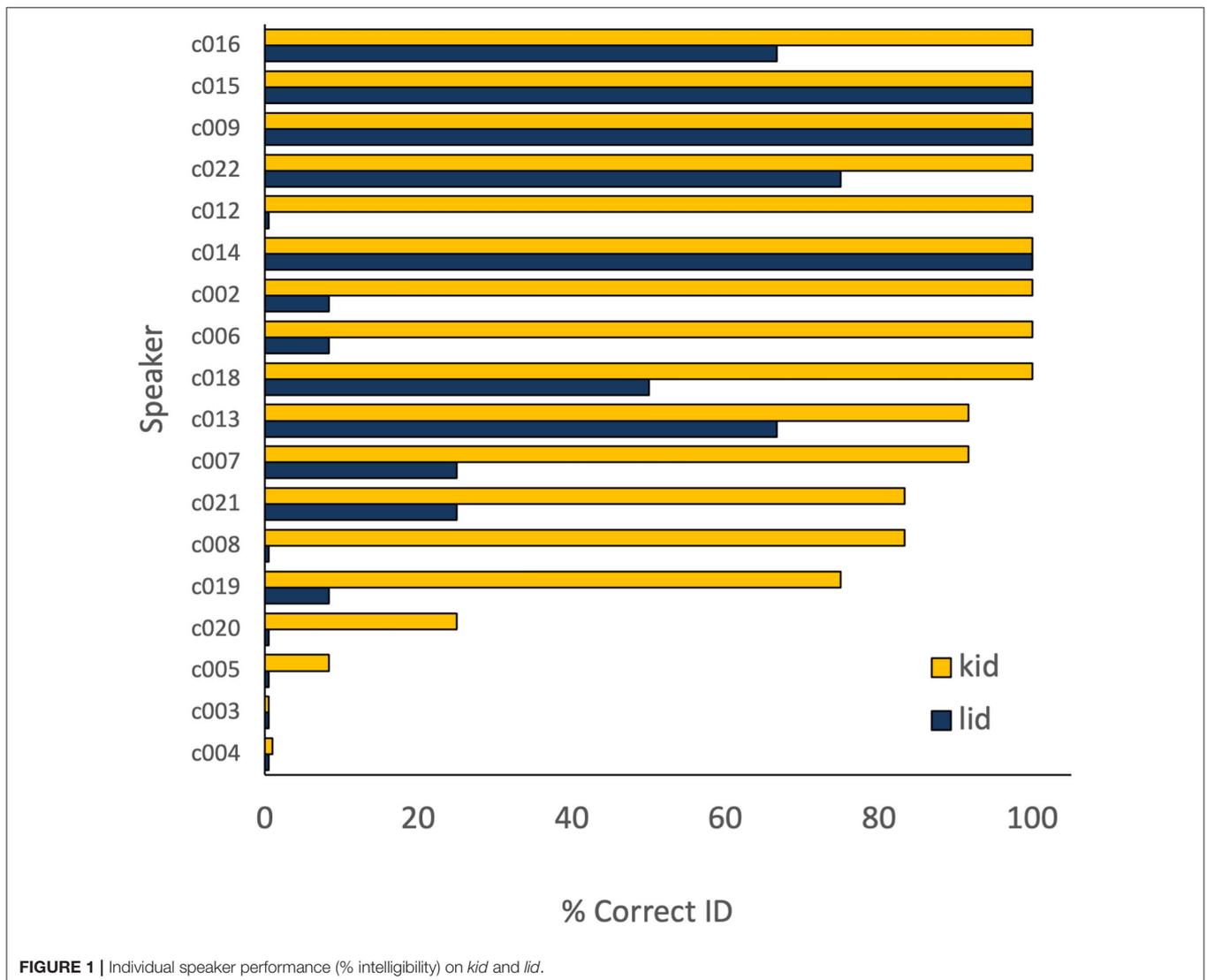
**TABLE 8** | Statistical comparisons of %-correct ID for words with identical rhymes.

Rhyme	Results*
it	[seat = feet] > heat
It	hit = sit
ik	speak > cheek
Ik	[sick = kick] > chick
ut	boot = suit
ŭt	put > foot
uk	Tuque > luke
ŭk	book = cook = look
id	feed = read
Id	kid > lid
ŭd	good = wood

\*No difference shown by "="; significant difference (Bonferroni  $p < 0.05$ ) shown by ">".

indicate significantly better performance on *kid*, and in fact, no speaker performed better on *lid*. However, several differences are evident across individual speakers. Three of them showed perfect performance on both words (c005, c009, and c014) and three showed 0% (or nearly 0%) on both (c003, c004, and c005). Discrepancies of 50 to 100 points between the two words occurred for seven speakers, and in those cases, the speaker performed better on *kid* than on *lid*.

No statistical difference was observed in performance on *hit* and *sit*. The individual speaker data shown in **Figure 2** reveal that eight speakers showed either identical or similar (within 10 points) correct ID scores for these words, suggesting comparable difficulty. However, the remaining data present a very different picture. Across all speakers, performance ranged from 0 to 100% for *sit* and from 0 to 83% for *hit*, suggesting slightly greater difficulty for the latter, while individual scores revealed a wide disparity in which word posed a greater challenge. Speaker c016 performed perfectly on *sit*, while scoring 0% on *hit*, with



**FIGURE 1** | Individual speaker performance (% intelligibility) on *kid* and *lid*.

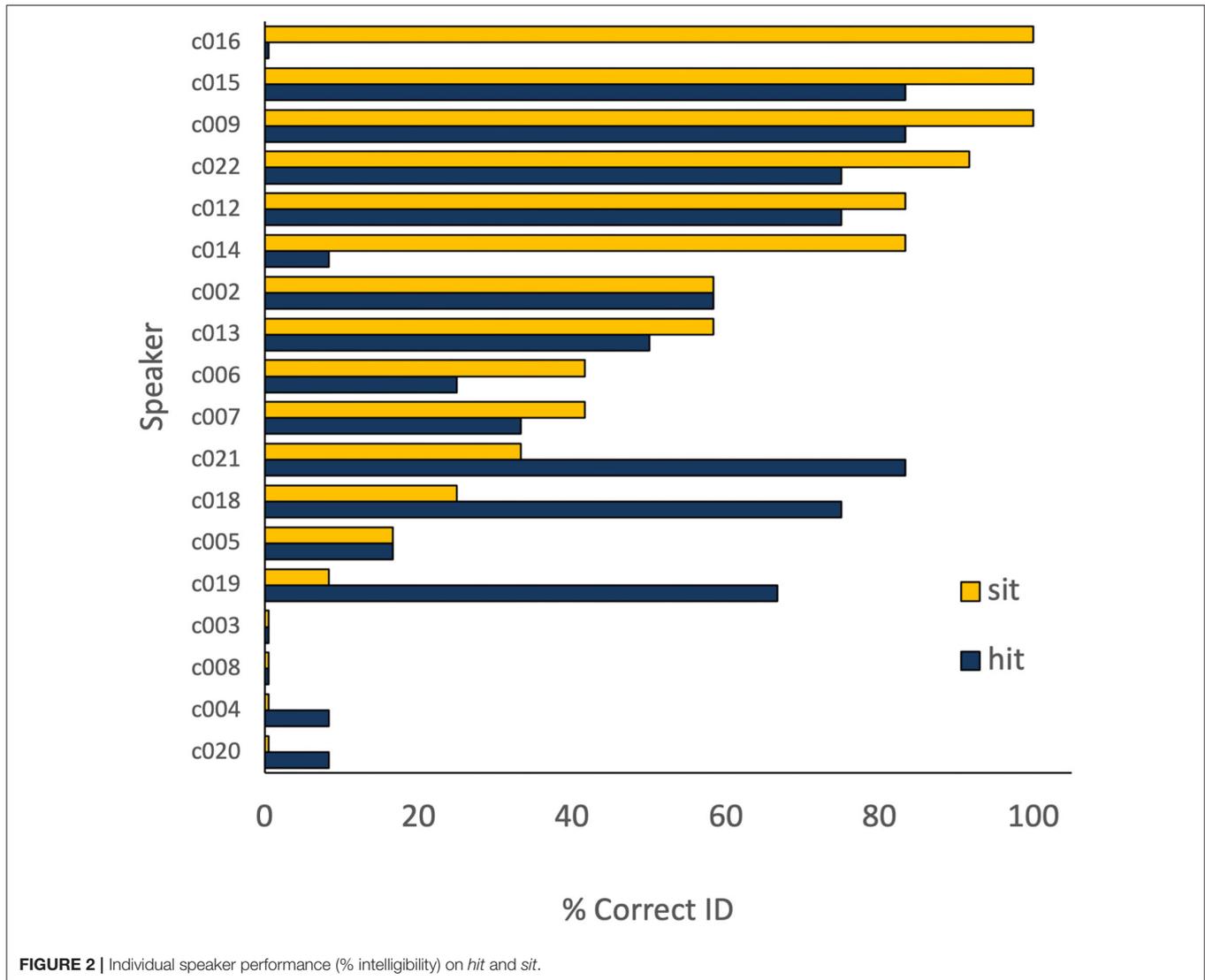


FIGURE 2 | Individual speaker performance (% intelligibility) on *hit* and *sit*.

speaker c014 also showing a sizeable discrepancy in the same direction. The pattern was reversed for speakers c021, c018, and c019, all of whom performed much better on *hit* than on *sit*. It appears, then, that the absence of a *post-hoc* statistical difference between the words was not the result of consistent performance across speakers. Rather, it was due to evaluating the average performance of speakers who showed opposite patterns of difficulty.

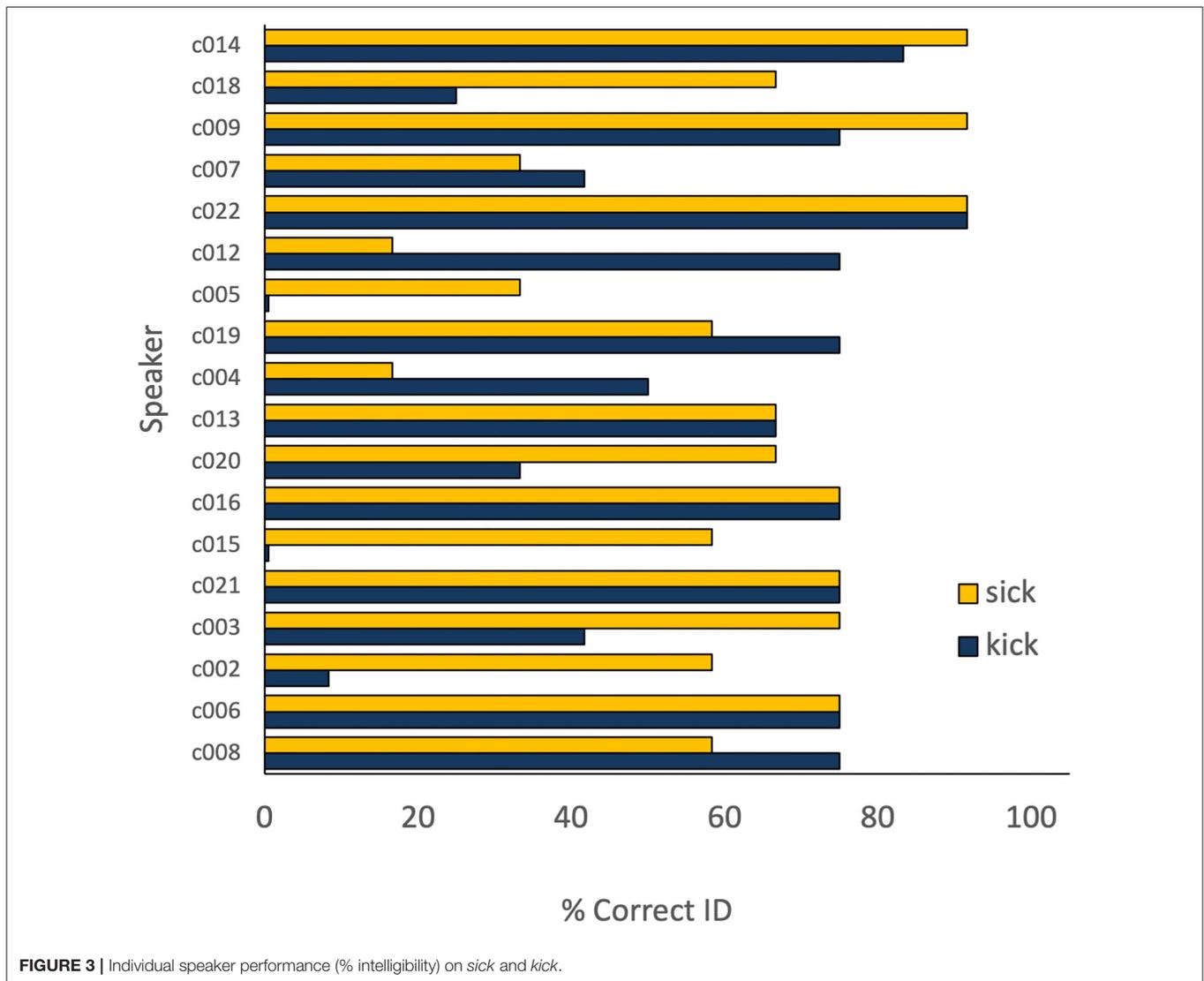
Similarly, for *sick* and *kick* no statistical difference was observed in vowel intelligibility. But as illustrated in **Figure 3**, that generalization does not accurately capture individual performance. Speakers c015, c018, and c002 all showed much better intelligibility for *sick*, while c012 and c014 performed much better on *kick*. Once again the lack of a statistical difference belies considerable variability among speakers.

Assuming a criterion of  $p < 0.05$ , none of the mixed-effects analyses yielded significant effects of either %USE or LOR. However, LOR just missed significance in all cases. For

that reason, the data were examined impressionistically for any general tendencies. With respect to rhymes, two of the top performers in **Table 3** (12 correct rhymes each) were tied for the longest LOR (6.8 years); however, a speaker with only slightly less time in Canada (6.7 years) was tied for poorest performance with two other speakers. Moreover, the speaker with the shortest LOR performed better than 12 of the other speakers. Daily use of English was also relatively high (72%) for the top performer (c009). However, c015 and c014, both of whom reported relatively low usage (16%, 33%), were the next two best performers.

### Vowel Confusions

Confusion patterns for the vowel productions are shown in **Table 9**. Nearly all non-target /i/ productions were classified by the judges as /i/, irrespective of coda consonant. However, while non-target /i/ was nearly always /i/ before /t/ and /d/, it was rarely so before /k/. Instead, target /ik/ was judged to contain



/e/ in 25% of cases, and some other vowel—most often /ɛ/—in another 14%. A parallel but larger discrepancy occurred for the back targets. Nearly all incorrect /u/ tokens were judged to be /ʊ/, and although non-target /ʊ/ items tended to be /u/ before /t/ and /d/, the majority of /ʊk/ targets (58%) were judged to contain /o/.

## DISCUSSION

The main question at issue in this study was whether or not it is feasible to specify a pedagogically useful hierarchy of difficulty for English close vowel production by Cantonese speakers of English. In this case, “difficulty” was operationalized in terms of intelligibility as assessed by a panel of four judges, and English tense-lax contrasts were selected for consideration because they are known to pose a challenge for Cantonese speakers due to different distributions in the two languages. “Pedagogically useful” refers to whether or not an empirically-based hierarchy could be expected to facilitate instruction by prioritizing certain

vowels or vowel + final consonant combinations for attention from teachers and learners.

### Difficulty by Vowel

From a statistical standpoint, a number of generalizations can be made to characterize the group performance observed in the study. First, open-syllable /i/ and /u/ were produced with nearly perfect intelligibility. Second, in checked syllables the following ordering emerged, with all differences being significant: /i/ > /u/ > /ʊ/ > /ɪ/. Thus, the two lax vowels were more difficult than the tense ones, and /ɪ/ proved most difficult of all, just as in Munro and Derwing’s (2008) study of Mandarin and Slavic language speakers.

### Difficulty by Rhyme

It was unsurprising that vowels in some rhymes were more difficult than in others, given Cantonese phonotactics. However,

**TABLE 9** | Confusions (%) by target rhyme.

Target Rhyme	Vowel identified (by judges)						
	i	I	e	u	ū	o	Other
it	83	14	2	0	0	0	1
It	46	50	2	0	0	0	3
ik	93	5	1	0	0	0	1
Ik	5	56	25	0	0	0	14
ut	0	1	0	73	18	2	6
ūt	0	0	0	27	72	0	1
uk	0	0	0	89	9	0	1
ūk	0	0	0	3	38	58	1
id	97	3	0	0	0	0	0
Id	40	60	0	0	0	0	0
ud	0	0	0	65	33	0	2
ūd	0	0	0	11	87	0	2

Shaded boxes show the largest ID category.

the patterns of difficulty did not fit a straightforward transfer-based explanation. In particular, vowels in “matching” rhymes (i.e., those that were roughly equivalent in both languages), did not exhibit higher rates of intelligibility. For instance, /ʊk/ (matching) exhibited much worse intelligibility than did /ʊt/ (non-matching), and both /ɪk/ (matching), and /ɪt/ (non-matching) showed similarly poor intelligibility overall. Nor could transfer be invoked to account for performance on vowels in Vd syllables, which do not occur in Cantonese. In fact, contrary to what a transfer account would suggest, these vowels were produced more intelligibly overall than the ones in matching Vt syllables.

The vowel confusion patterns (Table 9) indicated interesting asymmetries in performance. For the front targets, /ɪ/ was often produced as /i/ before /t/, but sometimes as /e/ (i.e., as lower than the target) before /k/. For back vowels, /ʊ/ was judged to be /u/ in most instances before /t/, but very often as /o/ (again lower) before /k/. The pattern of apparent tongue-lowering seen here brings to mind an observation made by Brière (1966) that a full understanding of transfer effects requires attention to relatively subtle phonetic details. In particular, as noted earlier, Zee (1991) reported relatively low tongue positions for the two lax vowels in Cantonese. It seems plausible that transfer of Cantonese articulatory patterns could lead to a perception on the part of the judges of lower-than-target English vowel categories. Nonetheless, this phonetic detail alone is insufficient to explain the different patterns between speakers. The reason why some speakers tended to produce a more /o/-like /ʊ/ while others did not must pertain to some aspect of individual learners’ phonological knowledge. An examination of these speakers’ L1 productions might prove useful in addressing this issue.

## Difficulty by Word

In contrast to the statistical significance of the rhyme effect, the wide-ranging word effect in these data was unexpected. Specifically, vowels in words sharing the same rhyme frequently

showed statistically different degrees of difficulty. This was true for each of the four vowels: for instance, /i/ in *heat* was produced less intelligibly than /i/ in *seat* or *feet*, /ɪ/ in *chick* was less intelligible than /ɪ/ in *sick* or *kick*, /u/ was less intelligible in *Luke* than in *tuque*, and /ʊ/ was less intelligible in *foot* than in *put*. It is possible to speculate about multiple reasons for this patterning, including effects of the initial consonant (as in Thomson, 2011), differential frequency of experience with the words, or prior encounters through borrowings of the same or similar words into Cantonese. None of these explanations can be verified as a causal factor in this study. More importantly, however, pursuing such lines of explanation may ultimately prove fruitless because of the complications of individual speaker performance described in the next section.

## Difficulty by Speaker

In spite of the group-based statistical tendencies discussed above, the most striking outcome of the study was the high degree of variability in performance across individual speakers. First, the between-speaker differences on different rhymes (Table 3) pose a serious problem for development of a detailed difficulty hierarchy. These differences do not correlate significantly with the socio-linguistic variables included in the study, i.e., use of English or length of residence in Canada. Nor can they be explained by an appeal to a “natural order of acquisition” for vowels in rhymes. If the latter applied, an implicational hierarchy should be visible within Table 3, such that success on some particular rhyme strongly predicts success on some statistically easier rhyme. However, indications of such patterning are exceedingly weak. Speaker c016, for instance, performed well on two of the very difficult rhymes (/ɪd/ and /ɪk/) but not on several that were statistically easier (e.g., /uk/, /ut/, /ud/). A similar point can be made about speakers c022 and c012. And speaker c014, who was the sole speaker to produce /ʊk/ rhymes at criterion, failed to produce an intelligible /ʊd/. Yet /ʊd/ was one of the easiest rhymes, having been produced intelligibly by 14 of the 18 participants.

The prognosis for establishing a meaningful hierarchy becomes considerably worse when one considers variability across words sharing the same rhyme. On the one hand, it might be proposed that Figure 1 illustrates a hierarchy of difficult at the level of the word. Not only was the *lid* vowel less intelligible overall than the *kid* vowel, but no speaker performed better on *lid* than on *kid*, and 13 speakers performed above the 80% criterion on *kid*. That outcome is consistent with a learning pattern in which the vowels in both words can eventually be acquired, but *kid* is mastered first. Because no longitudinal data were collected for this study, that account can be neither confirmed nor rejected. However, even if it proved true, other data do not appear explicable in terms of word-level hierarchies. Figure 2 is an especially compelling example of the Wade et al.’s (2020) observation that focussing on group-level performance can obscure important individual variability in speech data. Despite no statistical difference between the vowels in *sit* and *hit*, some speakers clearly performed much better on *sit*, while others did better on *hit*. The same type of concern can be raised about *sick* and *kick* in Figure 3.

## Linguistic and Sociolinguistic Predictors of a Hierarchy of Difficulty

The problems cited above are not the result of a lack of systematicity in the data. The mixed effects analyses in fact confirmed that broad generalizations could be made about the speakers' difficulties in terms of linguistic predictors: vowel (e.g., /ɪ/ was the most difficult of the four), rhyme (/ʊk/ was very difficult) and word (*lid* was especially difficult). Nonetheless, an analysis at a purely linguistic level could not possibly account for the quite idiosyncratic performance of the speakers in the study. Two sociolinguistic variables—length of Canadian residence and use of English—were included in the modeling, but neither contributed meaningfully to the outcome. One might propose including additional psycho-social variables (such as aptitude or motivation) to explain some of the individual differences. While doing so might improve the predictive power of the model, many aspects of the data could not possibly be accounted for in such a way.

In the results seen here, the most serious roadblock to understanding the production difficulties lies in the nearly opposite patterns of difficulty seen in the performance of individual speakers. For instance, even if it were found that differences in proficiency, aptitude or English use could explain why some speakers had difficulty with the vowel in *heat*, while others did not, one is still left to explain why some speakers found the *sit* vowel more difficult than the *hit* vowel and others showed the reverse ordering. It seems far-fetched to suppose that *any* quantitative dimension, such as amount of language use, could have differential effects on different speakers, making one word easier for some and a different word easier for others. Instead, the reasons for the differential performance appear to be idiosyncratic. Since the reasons might be tied to quality, rather than quantity, of language experience, understanding them might require examination of individual learners' experience with English in great detail to pinpoint the time at which particular lexical items were acquired, the interlocutors who modeled them and the speakers' frequency of exposure and use over months or years. Such a pursuit is out of the question in a retrospective study.

## Implications for Teaching

The central question posed in this study—whether or not a pedagogically useful hierarchy could be established for vowel intelligibility—must be answered mainly in the negative. In order to be useful to a classroom teacher, a difficulty hierarchy would need to pinpoint common difficulties that all or most learners in a given group of students would experience. Hypothetically, the most useful piece of information on relative difficulty would be that all, or virtually all, Cantonese learners of English have trouble with structure *x*, where *x* refers to a particular vowel, rhyme, or word. Such a finding might immediately suggest a focus of attention for the classroom, assuming that the structure is important for intelligibility. Next on the list of desiderata might be a finding that Cantonese speakers fall into perhaps two or three groups on the basis of whether they have trouble with structure *x*, structure *y* or structure *z*, perhaps determined

by their overall proficiency level. The teacher's strategy might then be to determine which category each student in the class belongs to, and to provide exercises tailored to the difficulties of each group of learners. It is obvious, however, that as the patterns of difficulty become more idiosyncratic (and therefore more complex), the less useful any hierarchy becomes. In this study, the degree of idiosyncrasy is high.

In terms of group patterns, a few generalizations can be offered from the present data, though their value is limited. For instance, /ɪ/ was the most difficult vowel. However, that difficulty has already been observed in other L1 groups, such as Mandarin and Slavic speakers (Munro and Derwing, 2008), and may even be a widespread difficulty for ESL learners in general. It is unlikely that teachers would be unaware of the common difficulties with English /ɪ/ after a modest amount of experience with learners. And despite the group-level performance on /ɪ/, its difficulty was inconsistent across rhymes and words, and highly idiosyncratic for individual learners. A useful message for teachers, then, is that a blanket strategy of "teaching /ɪ/" in all contexts to all learners would be inefficient.

The second most difficult vowel was /ʊ/, but here again idiosyncratic performance requires that the generalization be qualified. Looking more closely at a specific rhyme, the nearly universal difficulty with /ʊk/ seems to meet the criterion for a useful finding. All but one of the speakers in the study performed quite poorly on /ʊk/ items, producing this rhyme most often with /o/ and occasionally with /u/. However, an additional concern has to do with the functional load of the English /ʊ/-/oʊ/ and /u/-/oʊ/ distinctions, particularly in the /k/ context. In fact, English has only a handful of minimal pairs involving these vowels and they generally do not entail confusable words (e.g., *cook-Coke*, *took-toke*, *tuque-took*, *Luke-look*, *kook-cook*). The low frequency of many of these items further reduces the functional load of the contrasts. In short, mispronouncing /ʊk/ with one of the adjacent vowels would entail minimal costs in terms of intelligibility and /ʊk/ would rank low on a list of difficulties requiring attention in an English pronunciation class.

A useful recommendation that can be made on the basis of these data is for instructors to lower their expectations of L1-based error hierarchies and instead focus on identifying and addressing individual learner needs. In the context of an entire class of Cantonese ESL learners in Canada, it would probably be counterproductive to target the other rhymes or words found to be statistically difficult in this study. In each case, a sizeable proportion of the class could be expected to perform well, and time would be wasted on unneeded instruction. At the same time, items generally found to be easy, would not be easy for *all* learners, so failure to attend to "minority difficulties" would disadvantage some class members. Therefore, an individualized program tailored to individual needs is preferable to one based on group-based data.

The implementation of individualized instruction requires a suitable assessment tool to pinpoint problem sounds, rhymes and words with high functional loads and then provide learners with appropriate help. While more research is needed to determine how best to carry out pronunciation needs assessments, current trends in technology are helping to bring a needs-based approach

within the reach of instructors. Evidence shows that the benefits of training in L2 speech perception can transfer to production (Sakai and Moorman, 2018). Furthermore, identification of individual difficulties in perception can be achieved to some extent with currently available software, and computer-based perceptual instruction can be provided through high variability phonetic training (HVPT). The effectiveness of HVPT has been demonstrated in several studies (e.g., Thomson, 2012, 2018; Iino and Thomson, 2018). (For a discussion of its implementation for teaching purposes see Barriuso and Hayes-Harb, 2018). Also on the horizon are new applications of automatic speech recognition for pronunciation instruction (García et al., 2020) that offer a great deal of promise for individualized instruction that includes direct feedback on production.

This study was not designed to address the question of why performance on vowels is so idiosyncratic. In fact, this finding should interest theorists because of the challenge it poses to modeling that treats individual variability as “noise.” In this case, some findings do not seem explicable on the basis of language experience, aptitude, age of learning or any other quantitative variable because such dimensions are usually assumed to have uniform effects for all learners. Rather, idiosyncratic difficulties may arise because of very specific details of a language learner’s experience. It is possible, for instance, that a fossilized mispronunciation of particular word might result from encountering and using it in the very early stages of SLA when L2 perceptual and production processes are undeveloped. Yet the same mispronunciation might not occur in a rhyming word acquired later on. Perhaps further research will shed light on this issue. In the meantime, an advisable practice is to ensure that learners receive pronunciation instruction right from the beginning of the acquisition process.

In considering the above recommendations, it must be noted that the current work has a number of shortcomings, chief among them the limited number of speakers surveyed and the narrowness of the focus (i.e., four vowels). Obviously it cannot be assumed that the data are a complete representation of how Cantonese speakers in general perform on English vowels. Although the elicitation procedure was intended to encourage

speakers to produce representative exemplars of their typical pronunciation, it must be noted that a word production task cannot be assumed to capture the nuances that would appear in the full range of contexts in which speakers might communicate. In future work, investigators may wish to expand their focus to encompass productions of a wider range of speech sounds in a variety of speaking situations.

## DATA AVAILABILITY STATEMENT

The data sets presented in this study are available via email from the author: [mjmunro@sfu.ca](mailto:mjmunro@sfu.ca).

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by SFU Research Ethics Board, Simon Fraser University. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

## FUNDING

This research was funded by a grant from the Social Sciences and Humanities Research Council of Canada.

## ACKNOWLEDGMENTS

Preliminary findings from this study were presented at the Annual Meeting of the Canadian Acoustical Association in Vancouver in 2008. I thank Herman Li, Susan Morton, and Natasha Penner for their important contributions to data collection, and Ian Bercovitz and Barinder Thind for their invaluable work on the data analysis. Also appreciated are many discussions with my colleagues Tracey Derwing and Ron Thomson.

## REFERENCES

- Archibald, J. (2017). “Transfer, contrastive analysis and interlanguage phonology,” in *The Routledge Handbook of Contemporary English Pronunciation*, eds O. Kang, R. I. Thomson, and J. M. Murphy (London; New York, NY: Routledge), 9–24. doi: 10.4324/9781315145006-2
- Barriuso, T. A., and Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *CATESOL J.* 30, 177–194.
- Bohn, O.-S., and Garibaldi, C. (2017). “Production and perception of Danish front rounded/y: a comparison of ultimate attainment in native Spanish and native English speakers,” in *Romance-Germanic Bilingual Phonology*, eds M. Yavaş, M. Kehoe, and W. Cardoso (Sheffield: Equinox), 121–136.
- Brière, E. J. (1966). An investigation of phonological interference. *Language* 42, 768–796. doi: 10.2307/411832
- Brinton, D. M. (2018). Reconciling theory and practice. *CATESOL J.* 30, 283–300.
- Burgos, P., Cucchiari, C., van Hout, R., and Strik, H. (2014). Phonology acquisition in Spanish learners of Dutch: error patterns in pronunciation. *Lang. Sci.* 41, 129–142. doi: 10.1016/j.langsci.2013.08.015
- Carlisle, R. S. (1998). The acquisition of onsets in a markedness relationship: a longitudinal study. *Stud. Second Lang. Acquis.* 20, 245–260. doi: 10.1017/S027226319800206X
- Catford, J. C. (1987). “Phonetics and the teaching of pronunciation: a systemic description of English phonology,” in *Current Perspectives on Pronunciation: Practices Anchored in Theory*, TESOL, ed J. Morley (Washington, DC), 87–100.
- Chan, A. Y. W., and Li, D. C. S. (2000). English and Cantonese phonology in contrast: explaining Cantonese ESL learners’ English pronunciation problems. *Lang. Cult. Curri.* 13, 67–85. doi: 10.1080/07908310008666590
- Darcy, I., and Krüger, F. (2012). Vowel perception and production in Turkish children acquiring L2 German. *J. Phonet.* 40, 568–581. doi: 10.1016/j.wocn.2012.05.001

- Darcy, I., Park, H., and Yang, C.-L. (2015). Individual differences in L2 acquisition of English phonology: the relation between cognitive abilities and phonological processing. *Learn. Individ. Diff.* 40, 63–72. doi: 10.1016/j.lindif.2015.04.005
- Derwing, T. M., and Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. Amsterdam: Benjamins. doi: 10.1075/llt.42
- Derwing, T. M., Munro, M. J., and Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Lang. Learn.* 48, 393–410. doi: 10.1111/0023-8333.00047
- Derwing, T. M., and Rossiter, M. J. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Appl. Lang. Learn.* 13, 1–17.
- Eckman, F. R. (2004). From phonemic differences to constraint rankings: research on second language phonology. *Stud. Second Lang. Acquis.* 26, 513–549. doi: 10.1017/S027226310404001X
- Eckman, F. R., Elreyes, A., and Iverson, G. K. (2003). Some principles of second language phonology. *Second Lang. Res.* 19, 169–208. doi: 10.1191/0267658303sr2190a
- Edwards, J. G. H. (2017). “Pronunciation and individual differences,” in *The Routledge Handbook of Contemporary English Pronunciation*, eds O. Kang, R. I. Thomson, and J. M. Murphy (London; New York, NY: Routledge), 385–398. doi: 10.4324/9781315145006-24
- Flege, J. E. (1995). Second language speech learning: theory, findings, and problems. *Speech Percept. Linguist. Exp. Issues Cross Lang. Res.* 92, 233–277.
- Garcia, C., Nickolai, D., and Jones, L. (2020). Traditional versus ASR-based pronunciation instruction: an empirical study. *Calico J.* 37, 213–232. doi: 10.1558/cj.40379
- Gatbonton, E., Trofimovich, P., and Magid, M. (2005). Learners’ ethnic group affiliation and L2 pronunciation accuracy: a sociolinguistic investigation. *TESOL Q.* 39, 489–511. doi: 10.2307/3588491
- Hung, T. T. N. (2000). Towards a phonology of Hong Kong English. *World English.* 19, 337–356. doi: 10.1111/1467-971X.00183
- Iino, A., and Thomson, R. I. (2018). “Effects of web-based HVPT on EFL learners’ recognition and production of L2 sounds,” in *Future-Proof CALL: Language Learning as Exploration and Encounters-Short Papers From EUROCALL 2018*, eds P. Taalas, J. Jalkanen, L. Bradley, and S. Thouënsy (Research-publishing.net), 106–111. doi: 10.14705/rpnet.2018.26.821
- Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310
- Larson-Hall, J. (2015). *A Guide to Doing Statistics in Second Language Research Using SPSS and R*. Long Grove, IL: Routledge. doi: 10.4324/9781315775661
- Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *J. Second Lang. Pronunc.* 6, 310–328. doi: 10.1075/jslp.20050.lev
- Levis, J., and Cortes, V. (2008). “Minimal pairs in spoken corpora: Implications for pronunciation assessment and teaching,” in *Towards Adaptive CALL: Natural Language Processing for Diagnostic Language Assessment*, 197208.
- Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *Tesol Q.* 39, 369–377. doi: 10.2307/3588485
- Liu, D. (2021). Prosody transfer failure despite cross-language similarities: Evidence in favor of a complex dynamic system approach in pronunciation teaching. *J. Second Lang. Pronunc.* 7, 38–61. doi: 10.1075/jslp.18047.liu
- Meng, H., Lo, Y. Y., Wang, L., and Lau, W. Y. (2007). “Deriving salient learners’ mispronunciations from cross-language phonological comparisons,” in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)* (New York, NY: IEEE), 437–442. doi: 10.1109/ASRU.2007.4430152
- Moulton, W. G. (1962). Toward a classification of pronunciation errors. *Mod. Lang. J.* 46, 101–109. doi: 10.1111/j.1540-4781.1962.tb01773.x
- Munro, M. J. (2018). How well can we predict second language learners’ pronunciation difficulties? *CATESOL J.* 30, 267–281.
- Munro, M. J., and Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Lang. Learn.* 45, 73–97. doi: 10.1111/j.1467-1770.1995.tb00963.x
- Munro, M. J., and Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: an exploratory study. *System* 34, 520–531. doi: 10.1016/j.system.2006.09.004
- Munro, M. J., and Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: a longitudinal study of vowel production. *Lang. Learn.* 58, 479–502. doi: 10.1111/j.1467-9922.2008.00448.x
- Munro, M. J., and Derwing, T. M. (2011). The foundations of accent and intelligibility in pronunciation research. *Lang. Teach.* 44, 316–327. doi: 10.1017/S0261444811000103
- Munro, M. J., and Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *J. Second Lang. Pronunc.* 6, 283–309. doi: 10.1075/jslp.20038.mun
- Munro, M. J., Derwing, T. M., and Thomson, R. I. (2015). Setting segmental priorities for English learners: evidence from a longitudinal study. *Int. Rev. Appl. Linguist. Lang. Teach.* 53, 39–60. doi: 10.1515/iral-2015-0002
- Munro, M. J., Flege, J. E., and MacKay, I. R. A. (1996). The effects of age of second language learning on the production of English vowels. *Appl. Psycholinguist.* 17, 313–334. doi: 10.1017/S0142716400007967
- Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: investigating motivation as a time-varying predictor of pronunciation development. *Mod. Lang. J.* 102, 199–217. doi: 10.1111/modl.12461
- Nagle, C. L., and Huensch, A. (2020). Expanding the scope of L2 intelligibility research: intelligibility, comprehensibility, and accentedness in L2 Spanish. *J. Second Lang. Pronunc.* 6, 329–351. doi: 10.1075/jslp.20009.nag
- Nilsen, D. L. F., and Nilsen, A. P. (2010). *Pronunciation Contrasts in ENGLISH*. Amsterdam: Waveland Press.
- Nimz, K., and Khattab, G. (2020). On the role of orthography in L2 vowel production: the case of Polish learners of German. *Second Lang. Res.* 36, 623–652. doi: 10.1177/0267658319828424
- O’Brien, M. G., and Smith, C. (2010). Role of first language dialect in the production of second language German vowels. *Int. Rev. Appl. Linguist. Lang. Teach.* 48, 297–330. doi: 10.1515/iral.2010.013
- Okuno, T., and Hardison, D., M. (2016). Perception-production link in L2 Japanese vowel duration: training with technology. *Lang. Learn. Technol.* 20, 61–80.
- Saito, K., Dewaele, J.-M., and Hanzawa, K. (2017). A longitudinal investigation of the relationship between motivation and late second language speech learning in classroom settings. *Lang. Speech* 60, 614–632. doi: 10.1177/0023830916687793
- Sakai, M., and Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Appl. Psycholingu.* 39, 187–224. doi: 10.1017/S0142716417000418
- Sewell, A. (2017). Functional load revisited: reinterpreting the findings of ‘lingua franca’ intelligibility studies. *J. Second Lang. Pronunc.* 3, 57–79. doi: 10.1075/jslp.3.1.03sew
- Swan, M., and Smith, B. (2001). *Learner English: A Teacher’s Guide to Interference and Other Problems*. Cambridge: CUP. doi: 10.1017/CBO9780511667121
- Sweet, H. (1900). *The Practical Study of Languages: A Guide for Teachers and Learners*. Glasgow: H. Holt.
- Sypiańska, J. (2016). Multilingual acquisition of vowels in L1 Polish, L2 Danish and L3 English. *Int. J. Multilingual.* 13, 476–495. doi: 10.1080/14790718.2016.1217606
- Thomson, R. I. (2011). Computer assisted pronunciation training: targeting second language vowel perception improves pronunciation. *Calico J.* 28, 744–765. doi: 10.11139/cj.28.3.744-765
- Thomson, R. I. (2012). Improving L2 listeners’ perception of English vowels: a computer-mediated approach. *Lang. Learn.* 62, 1231–1258. doi: 10.1111/j.1467-9922.2012.00724.x
- Thomson, R. I. (2018). High variability [pronunciation] training (HVPT) A proven technique about which every language teacher and learner ought to know. *J. Second Lang. Pronunc.* 4, 208–231. doi: 10.1075/jslp.17038.tho
- Thomson, R. I., and Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: a narrative review. *Appl. Linguist.* 36, 326–344. doi: 10.1093/applin/amu076
- Wade, L., Lai, W., and Tamminga, M. (2020). The reliability of individual differences in VOT imitation. *Lang. Speech* 23830920947769. doi: 10.1177/0023830920947769
- Walz, J. (1980). An empirical study of pronunciation errors in French. *French Rev.* 53, 424–432.

- Wardhaugh, R. (1970). The contrastive analysis hypothesis. *TESOL Q.* 4, 123–130. doi: 10.2307/3586182
- White, L. (2018). “What is easy and what is hard,” in *Meaning and Structure in Second Language Acquisition: In honor of Roumyana Slabakova*, eds J. Cho, M. Iverson, T. Judy, T. Leal, and E. Shimanskaya (Benjamins), 263–282. doi: 10.1075/sibil.55.10whi
- Wong, J. W. S. (2013). “The effects of perceptual and/or productive training on the perception and production of English vowels /I/ and /i:/ by Cantonese ESL learners,” in *Interspeech*, 2113–2117.
- Wong, J. W. S. (2015). “Comparing the perceptual training effects on the perception and production of English high-front and high-back vowel contrasts by Cantonese ESL learners,” in *Proceedings of the 18th International Congress of Phonetic Sciences*, ed The Scottish Consortium for ICPhS (University of Glasgow).
- Zee, E. (1991). Chinese (Hong Kong Cantonese). *J. Int. Phonet. Assoc.* 21, 46–48. doi: 10.1017/S0025100300006058

**Conflict of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher’s Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

*Copyright © 2021 Munro. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.*