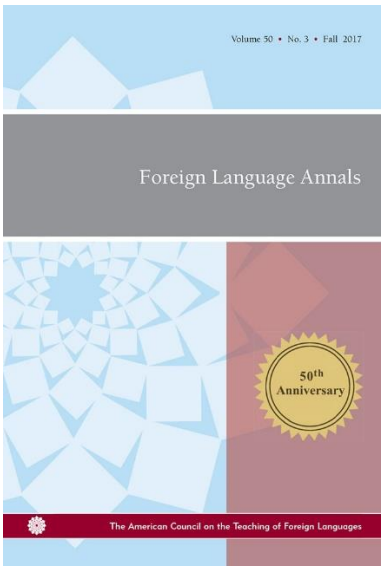

Effects of benchmarking and peer-assessment on French learners' self-assessments of accentedness, comprehensibility, and fluency

Aki Tsunemoto, Pavel Trofimovich, Josée Blanchet, Juliane Bertrand, and Sara Kennedy

Tsunemoto, A., Trofimovich, P., Blanchet, J., Bertrand, J., & Kennedy, S. (2022). Effects of benchmarking and peer-assessment on French learners' self-assessments of accentedness, comprehensibility, and fluency. *Foreign Language Annals*, *55*(1), 135–154. https://doi.org/10.1111/flan.12571

Abstract

This study examined the effect of benchmarking and peer-assessment activities on second language (L2) French learners' self-assessments of accentedness, comprehensibility, and fluency. The learners, who included 25 L2 French students enrolled in a 15-week university-level French course, recorded two oral presentations at the beginning and the end of the course and self-assessed their accentedness, comprehensibility, and fluency four times. In addition to regular course instruction, the treatment group also engaged in benchmarking (discussing and applying pre-established evaluation criteria) and peer-assessment (evaluating peers' speaking performances). The students' self-assessments were compared with ratings of accentedness, comprehensibility, and fluency by 10 native-speaking French raters. At the end of the course (i.e., for the second oral presentation), the treatment group showed greater alignment in self-assessment of comprehensibility than the comparison group, relative to the external raters' assessments. Results highlight the value of assessment-focused activities targeting L2 learners' awareness of pronunciation.

Keywords: self-assessment; L2 French; pronunciation; comprehensibility; university teachers

The Challenge

Accurate self-assessment is a key skill characterizing self-regulated, autonomous language students. However, many students find it particularly challenging to self-assess their second language pronunciation. What activities can second language teachers of French implement to help their students become accurate at evaluating how accented, comprehensible, and fluent they are in French?

Effects of Benchmarking and Peer-Assessment on French Learners' Self-Assessments of Accentedness, Comprehensibility, and Fluency

Introduction

Self-assessment—the metacognitive skill of evaluating one's own performance—has become a popular topic with researchers and practitioners in second language (L2) teaching and learning (Li & Zhang, 2021). Featured in educational policy statements (Butler & Lee, 2010) and included in formative and achievement tests (ACTFL, 2012; Babaii et al., 2016; Hung, 2019), self-assessment is considered central to learners' autonomy. Learners who can accurately evaluate their language skills can relate their performance to instruction (Chen, 2008; Peirce et al., 1993), which ideally promotes instruction-relevant behaviors, such as seeking additional language practice and continuing with language study.

However, compared to self-assessment in other language skills (Denies & Janssen, 2016; Matsuno, 2009; Suzuki, 2015), self-assessment in L2 pronunciation has received relatively little attention. Accurate self-assessment in pronunciation is particularly important because teachers often lack time, resources, and training to provide pronunciation-specific instruction (Georgiou, 2019; Murphy, 2014). Unlike other skills, such as listening and reading (Li & Zhang, 2021), pronunciation is also difficult for learners to self-assess, especially when it comes to its global dimensions (Trofimovich et al., 2016), including accentedness (how closely the learner approximates the target language variety), comprehensibility (how easy the learner is to understand), and fluency (how fluid the learner's speech sounds). Learners generally have little awareness of their own challenges, unable to articulate what makes their speech accented, difficult to understand, or disfluent (Derwing, 2003; Strachan et al., 2019).

One notable absence in prior work is research examining the effectiveness of pedagogical interventions targeting L2 learners' self-assessment. Familiarizing learners with assessment criteria and engaging them in peer-assessment appear beneficial for self-assessment (Babaii et al., 2016; Chen, 2008). However, this work has largely targeted only specific aspects of pronunciation, such as segmental errors, rather than the global dimensions of accentedness, comprehensibility, and fluency that are central to how interlocutors assess L2 learners (Derwing & Munro, 2015). In addition, most prior work has included short-term interventions, often conducted outside language instruction, and has disproportionately targeted L2 English, with only a handful of studies focusing on other languages. Therefore, in this 15-week classroom-based study, we examined the effectiveness of benchmarking and peer-assessment as pedagogical activities aimed at improving the accuracy of L2 French learners' self-assessment of accentedness, comprehensibility, and fluency.

Background Literature

Self-Assessment and L2 (French) Pronunciation

Prior work in languages other than French (mostly English) has shown that L2 learners' self-assessment in speaking tasks often deviates greatly from teachers' evaluations of the same tasks (Lee & Chang, 2005; Patri, 2002). However, training in self-reflection and self-assessment appears to lead to more convergence between learners' self-assessments and teachers' evaluations (Babaii et al., 2016; Chen, 2008; see also Meritan & Mroz, 2019). Similarly, repeated self-assessment seems to impact learners' language performance positively, such that those who self-assess their speech over time improve in various dimensions of speech, including fluency, pronunciation, and connected speech processes (de Saint Leger, 2009; Kissling & O'Donnell, 2015), likely through increased awareness of these dimensions (for empirical work

linking L2 learners' pronunciation development to measures of self-reflection and awareness, see Meritan, 2020; Meritan & Mroz, 2019; Sturm, 2013, 2019).

Emerging evidence from L2 French similarly suggests that instruction impacts learners' self-assessment positively. For example, Dolosic et al. (2016) reported greater alignment in self-assessment for learners enrolled in a 4-week intensive summer language program. Early in the program, there was no association between measures of learners' speaking (e.g., articulation rate) and their self-assessments (Can-Do statements targeting everyday tasks in French). However, by the program's end, such associations emerged, likely because learners developed awareness of at least some aspects of their speaking performance. Similarly, L2 French learners receiving phonetic training in a French pronunciation course revealed moderate-to-strong correlations between their self-assessments in a read-aloud task and the evaluations of the same task by two expert raters (Lappin-Fortin & Rye, 2014). Although learners' self-assessment did not fully align with external ratings, post-instruction self-assessments were more aligned than pre-instruction self-assessments for several specific pronunciation targets (e.g., liaisons, silent *e*).

Self-Assessment of Global Dimensions of Pronunciation

While it is reassuring to know that L2 learners' self-assessments of specific aspects of pronunciation might align with external evaluations, there appears to be low agreement between learners' self-assessments and external evaluations of global dimensions of L2 speech, such as accentedness, comprehensibility, and fluency. Accentedness has most recently been defined as "[t]he extent of differences perceived by speakers of one linguistic variety when listening to speakers of other varieties" (Derwing & Munro, 2015, p. 175) whereas comprehensibility refers to "[t]he degree of effort required by a listener to understand an utterance" (Derwing & Munro, 2015, p. 176). Many researchers and teachers have been promoting intelligible, comprehensible

L2 speech as the main goal of L2 teaching and learning (e.g., Levis, 2020), which is contrasted with a focus on non-accented performance whereby L2 learners could pass for native speakers of the target language variety. When it comes to L2 speech that is clearly understandable to the listener, intelligibility is by far the more informative dimension (as it captures the extent to which listeners actually understand L2 speech); however, comprehensibility is a useful alternative, in that it is an intuitive, reliable, and easy-to-use scalar measure of understanding (Kennedy & Trofimovich, 2019; Nagle, 2019). Nevertheless, many L2 speakers (and their teachers) aspire to reach non-accented, nativelike pronunciation (Derwing, 2003; Inceoglu, 2019), which means that accentedness is also a key dimension relevant to learners and their teachers, even though L2 instruction might be communicative in nature and it might focus on the attainment of intelligible, comprehensible L2 speech. Finally, the dimension of fluency, which captures "[t]he degree to which speech flows easily without pauses and other dysfluency markers such as false starts" (Derwing & Munro, 2015, p. 177), is an important component of L2 speech because it contributes to listener engagement (Derwing et al., 2004). Often targeted in high- and low-stakes assessments (Derwing & Munro, 2015), accentedness, comprehensibility, and fluency thus capture the essential global aspects of L2 learners' pronunciation performance.

To date, only a handful of studies have focused on learners' self-assessment of global dimensions of L2 speech, and such studies have revealed discrepancies between self- and other-assessments. For example, Mitterer et al. (2020) investigated L2 speakers' bias in rating their own versus fellow L2 speakers' accentedness. In this study, 24 female learners of L2 English consistently rated their own productions, which were acoustically altered to sound like male speakers, as being less accented than the productions of other (male-sounding) learners, without realizing that the preferred voice actually belonged to them. Put differently, these L2 speakers

overestimated their own accent (judging themselves as less accented), relative to the assessments by fellow learners. In another L2 English study, Trofimovich et al. (2016) examined how closely learners' judgments of accentedness and comprehensibility matched the evaluations by external raters. Correlational analyses across 134 university-level L2 speakers revealed no meaningful associations between self- and other-assessments for accentedness and only a weak relationship for comprehensibility, implying that even high-level L2 speakers have difficulty understanding how accented and comprehensible they sound to external listeners. Similarly, Préfontaine (2013) found only a moderate relationship between self- and other-assessments of fluency for 40 learners of L2 French in a university immersion context, again suggesting that learners' own fluency judgments do not align with the evaluations by expert raters. Thus, accentedness, comprehensibility, and fluency, with multiple linguistic cues (e.g., individual sounds, prosody, pausing, speech rate, lexical sophistication) contributing to each dimension (S. Suzuki & Kormos, 2020; Trofimovich & Isaacs, 2012), might be particularly complex for learners to self-assess in the absence of targeted instruction.

Instructional Interventions Targeting Self-Assessment

Although repeated self-assessment, especially paired with instruction focusing on specific aspects of pronunciation, might enhance L2 learners' self-assessment accuracy (Dolosic et al., 2016; Kissling & O'Donnell, 2015; Lappin-Fortin & Rye, 2014), accurate self-assessment of global dimensions of L2 pronunciation might crucially depend on targeted interventions. Two relevant interventions attested across educational psychology and language assessment are benchmarking and peer-assessment (Dunning et al., 2004; Elder et al., 2007; Peirce et al., 1993). Benchmarking refers to pedagogical activities where teachers provide learners with exemplars of work demonstrating a range of performance levels (low to high) so that learners first engage in a

discussion of possible assessment standards and criteria (which are either generated by learners themselves or are provided for discussion by the teacher) and then practice using these pre-established standards and criteria in their own assessments (Hendry et al., 2009). For instance, Babaii et al. (2016) showed that setting performance benchmarks for learners and sharing rating criteria with them led to improved alignment in self- and other-assessment. In that study, the instructor developed rating criteria for various dimensions of speaking (e.g., fluency, grammar, vocabulary, confidence, organization). After learners practiced using these criteria and assessed three sample presentations, they self-assessed their own speaking performances, demonstrating greater alignment between their own and external raters' evaluations.

Peer-assessment, which refers to the practice of learners assessing the performances of fellow L2 speakers, contributes to learners' "greater understanding of the nature and process of assessment" (Hansen Edwards, 2013, p. 734). Arguably, peer-assessment enables learners to make critical comparisons between peers' performances and the target model (Topping, 1998), enhancing their understanding of performance benchmarks. For example, working with L2 English university learners in a Taiwanese university, Chen (2008) had L2 learners evaluate their own and their peers' oral presentations for content, language, delivery, and manner. At first, learners' self-assessments diverged from teacher-assessments. However, after engaging in repeated peer-assessment and receiving feedback from peers and the teacher over time, learners narrowed the gap between their self-assessments and teacher's evaluations of their performances. Thus, both benchmarking and peer-assessment appear beneficial in helping L2 learners internalize assessment standards and consider their performance in relation to that of peers.

The Present Study

Although there is a growing body of work focusing on promoting L2 learners' self-assessment in

L2 pronunciation, most past research has targeted learners of L2 English, generally in cross-sectional (one-shot) designs (Mitterer et al., 2020; Trofimovich et al., 2016). The few longitudinal studies conducted in L2 French are promising, in that they suggest that instruction positively shapes self-assessment accuracy. Nevertheless, these findings remain tentative, because learners in these studies evaluated their speech using general Can-Do statements, rather than pronunciation-specific criteria (Dolosic et al., 2016), or used different self-assessment rubrics pre- and post-instruction (Lappin-Fortin & Rye, 2014), making comparisons difficult. Moreover, there are presently few investigations of specific instructional interventions targeting self-assessment, and those that exist often deal with L2 speaking, not pronunciation, mostly in L2 English (Babaii et al., 2016; Chen, 2008). This work also has methodological limitations, such as the absence of comparison groups, which makes it hard to attribute improved self-assessment accuracy to a specific intervention rather than instruction more generally. Finally, to the best of our knowledge, there is no work centering on three global dimensions of pronunciation—accentedness, comprehensibility, and fluency—through targeted self-assessment interventions. As discussed previously, these global dimensions are critical to the impressions that L2 learners make on their interlocutors (e.g., how accented or fluent learners sound or how difficult they are to understand). These dimensions are also important to learners themselves, as many learners wish to speak fluently, comprehensibly, and with little foreign accent (Derwing, 2003; Inceoglu, 2019). Therefore, in the present study, we examined the effects of two pedagogical interventions—benchmarking and peer-assessment—on the extent to which learners' self-assessment of L2 French accentedness, comprehensibility, and fluency align with assessment by external raters. We targeted L2 French learners enrolled in a 15-week French oral communication course at a French-medium university, where learners in the treatment group

were exposed to benchmarking and peer-assessment activities, in addition to regular instruction, whereas those in the comparison group engaged in regular instruction only. This study was guided by the following research question: To what extent do benchmarking and peer-assessment activities impact L2 French learners' self-assessment of accentedness, comprehensibility, and fluency, relative to evaluations by external raters?

Method

Students

The initial participant sample included 45 L2 French adult students from two intermediate-level sections of a listening and speaking course at a French-medium university in Quebec, Canada. Because of data attrition during the 15-week term, from cases when individual students were absent during one or more of the five data collection events or missed one or more classes featuring benchmarking or peer-assessment activities (see below), we retained only the data for the 25 students who provided all the datapoints in this data-heavy project. Thus, the final dataset, which included 15 students in the treatment group and 10 students in the comparison group, was the strongest possible, allowing direct comparisons of the same students' performances. The students, whose background characteristics are summarized in Table 1, represented nine different ethnolinguistic groups, the largest being Spanish (9), Mandarin (4), Arabic (2), and Persian (2). In terms of their L2 French skills, all students were considered intermediate to advanced in proficiency because they either scored at a lower C1 level on the language program's in-house test prior to enrolling in the target course or had passed a previous French language course at a B2 level. Independent-samples *t* tests comparing the students' background characteristics across the two groups (see Table 1) revealed no statistically significant differences, $t < -1.27$, $p > .218$.

Table 1 *Descriptive Statistics for Students*

Background variables	Treatment (n = 15)			Comparison (n = 10)		
	M	*SD*	Range	*M*	*SD*	Range
Age (years)	32.2	6.0	21–40	34.4	8.8	21–51
Length of residence in Quebec (years)	3.5	3.6	0–13	3.6	3.2	0–10
Age French instruction began (years)	25.2	9.5	5–37	26.7	10.6	7–40
Length of French instruction (years)	0.9	0.5	0–2	0.1	0.3	0–1
Self-rated French speaking (1–9 scale)	5.7	1.3	3–8	5.6	1.3	4–7
Self-rated French listening (1–9 scale)	6.3	1.5	2–8	6.3	1.8	3–9
Use of French at school (0–100%)	3.2	1.7	1–7	2.3	1.1	1–4
Use of French at home (0–100%)	3.6	6.1	0–20	1.7	1.8	0–6
Use of French socially (0–100%)	2.0	2.3	0–8	1.5	2.3	0–8

Course

All students were enrolled in a 15-week French listening and speaking course that met once per week for three hours, with one hour spent in a multimedia lab. The course, which targeted the development of intelligible pronunciation in speaking and successful decoding of connected speech in listening, included form-focused instruction on individual segments (sounds), especially those that are specific to the Quebec variety of French (e.g., diphthongization), and on prosody, with particular attention devoted to connected speech processes, such as liaison and enchainement. The treatment and the comparison groups were taught by the same instructor, a native speaker of Quebec French, with graduate degrees in linguistics and over 10 years of experience teaching L2 French at a postsecondary level. Another instructor, who had developed the course, worked in close collaboration with the course instructor, ensuring that all activities

were identical across the two sections. Typical class activities included oral presentations, dictation, listening to authentic recorded materials (e.g., dialogues), and predicting pronunciation from spelling, all of which provided ample opportunities for students to listen to and produce French while their attention was drawn to intelligible pronunciation.

Quasi-Experimental Design

During the 15-week course, all students delivered two oral presentations (both included in the course curriculum), which they also self-assessed for accentedness, comprehensibility, fluency. However, only the treatment group engaged in benchmarking and peer-assessment activities. As shown in Figure 1 illustrating the study's design, the students recorded Presentation 1 individually in a multimedia lab in Week 3. The presentation was based on a video related to the course reading about the arrival of refugees in Quebec in the 1970s. After reading and discussing the article in class, the students were instructed to describe and respond to the content of the reading in an approximately 2-minute recording. Although the students could re-record their speech, the pausing function was disabled so they had the (rarely chosen) option to narrate the entire presentation again. The students self-assessed Presentation 1 on three occasions: in Week 3 (immediately after the presentation was recorded) and then again in Week 6 (after benchmarking activities) and Week 12 (after peer-assessment activities). For Presentation 2, the students chose an individual topic that they considered socially or politically significant to Quebec (e.g., access to education, indigenous population) for an approximately 10-minute presentation, which they practiced in class and at home for about a month. Presentation 2 was delivered in class on a pre-assigned day during Weeks 7–10 and was audio-recorded, with several students presenting on a given day. The students self-assessed Presentation 2 in Week 12 (after benchmarking and peer-assessment activities).

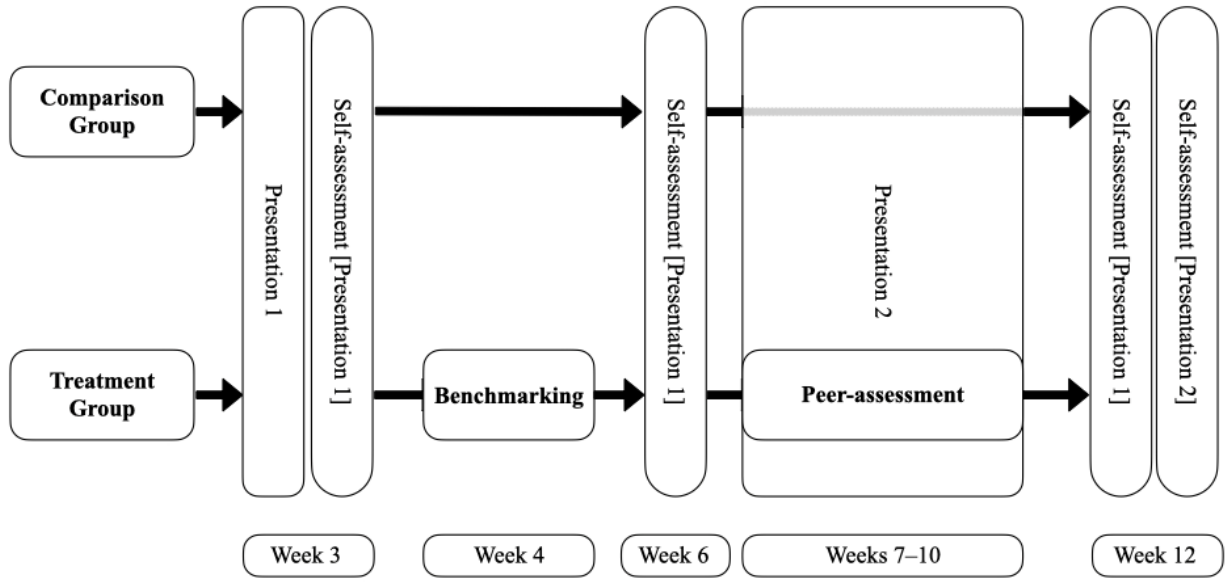

Figure 1. Quasi-experimental design.

Target Interventions

As shown in Figure 1, the target interventions were implemented only in the treatment group. The benchmarking activity took place in Week 4, where the teacher first performed a sample 10-minute presentation to model in-class Presentation 2 for both comparison and treatment groups. All students then received a brief summary of the teacher's expectations for this assignment, after which the comparison class engaged in a 30-minute small-group and whole-class discussion of possible topics, presentation styles, and study approaches for Presentation 2. In contrast, the treatment class performed a 30-minute benchmarking activity in groups of three, where the teacher first introduced the concept of successful L2 speech performance (i.e., speech that is comprehensible and fluent and that is not heavily coloured by foreign accent), and the students then rated short excerpts of six sample presentations (i.e., Presentation 2) by L2 speakers enrolled in the same course in a previous term (1–2 minutes in length), with speaker identities

unknown to the students. The sample presentations were purposefully chosen by the instructor who had designed the course to include various (low to high) levels of speaking performances, with varying levels of segment- and prosody-level pronunciation as well as fluency issues. These issues included both weaknesses (e.g., consonant omissions or substitutions, frequent but short hesitations, flat or incorrect intonation patterns) and strengths (e.g., accurate articulation of segments, fluid delivery, natural-sounding prosody), which allowed learners to discuss various linguistic features in each speech sample. For benchmarking, the students first discussed the target scales, printed in a booklet as 100-millimeter lines, with only anchor points labeled (see below). They then listened to each recording (as many times as needed by groups or individual students) and performed the ratings individually. Finally, they engaged in a group discussion of their motivations for each speaker's ratings to arrive at an agreed-upon description of each rated construct (e.g., in terms of linguistic features relevant to each). Although the teacher acted as a moderator, she provided no specific feedback or input.

The peer-assessment activity took part during eight course meetings (twice per week) in Weeks 7–10, when the students delivered Presentation 2 in front of classmates (with 3–5 students presenting per day). The students in the treatment group evaluated each presenter using the same scales for accentedness, comprehensibility, and fluency after each in-class (live) presentation, with the scales printed in a booklet following the same format as in the benchmarking activity. Thus, each in-class presentation was evaluated by each peer, such that each of the 15 treatment group's students evaluated the remaining 14 students' presentations. Neither the teacher nor peers gave feedback on peer-assessments, with the students engaged in peer-assessment individually. Also, the students did not have access to peers' evaluations, which could have impacted their self-assessments. The students in the comparison group provided no

ratings of peer presentations. However, to provide a comparable level of engagement during presentations, they were instructed to attend to new content or vocabulary, practice efficient listening and note-taking skills, and engage in otherwise good pragmatic behaviors (e.g., providing non-verbal backchannels, signalling interest). The students in the treatment group were aware that the benchmarking and peer-assessment activities (and self-assessments of Presentation 1 and 2) were unrelated to course performance and course grades.

Student Self-Assessments

All student assessments (self-assessment, benchmarking, peer-assessment) were carried out using identical scales (see Figure 2), with only anchor points labeled for accentedness (1 = *accent marqué* [strong accent], 100 = *aucun accent* [no accent]), comprehensibility (1 = *impossible à comprendre* [impossible to understand], 100 = *facile à comprendre* [easy to understand]), and fluency (1 = *trop rapide ou trop lent* [too fast or too slow], 100 = *naturel/optimal* [natural/optimal]). The constructs were defined for students at each rating episode. Accentedness was described as the extent to which the speech differed from a production pattern expected of a native speaker of (Quebec) French, with "strong accent" describing the speech that departs heavily from a native speaker's production and "no accent" referring to nativelike production. This definition of accentedness is compatible with Derwing and Munro's (2015) conceptualization of this construct, in that it captures listener perceptions of how a given speech sample differs from the expected or target speech variety. Comprehensibility was introduced as the degree of effort required by the listener to understand the speech. Fluency concerned the pace of speech, with fluent performance characterized by few pauses and hesitation and an optimal speaking rate (not too slow and not too fast). For repeated self-assessments, the students in both treatment and comparison groups evaluated their Presentation 1

performance immediately after recording and listening to it in the lab in Week 3; they assessed their Presentation 1 again in Weeks 6 and 12, and evaluated their Presentation 2 in Week 12 (see Figure 1) by listening to a 45-second excerpt of their speech (taken from the beginning of each presentation) played to them individually in a multimedia lab. Because the students may have had limited experience with self-assessment prior to the course, the instructors introduced self-assessment to students as a valuable activity that could raise students' awareness of their pronunciation strengths and challenges.

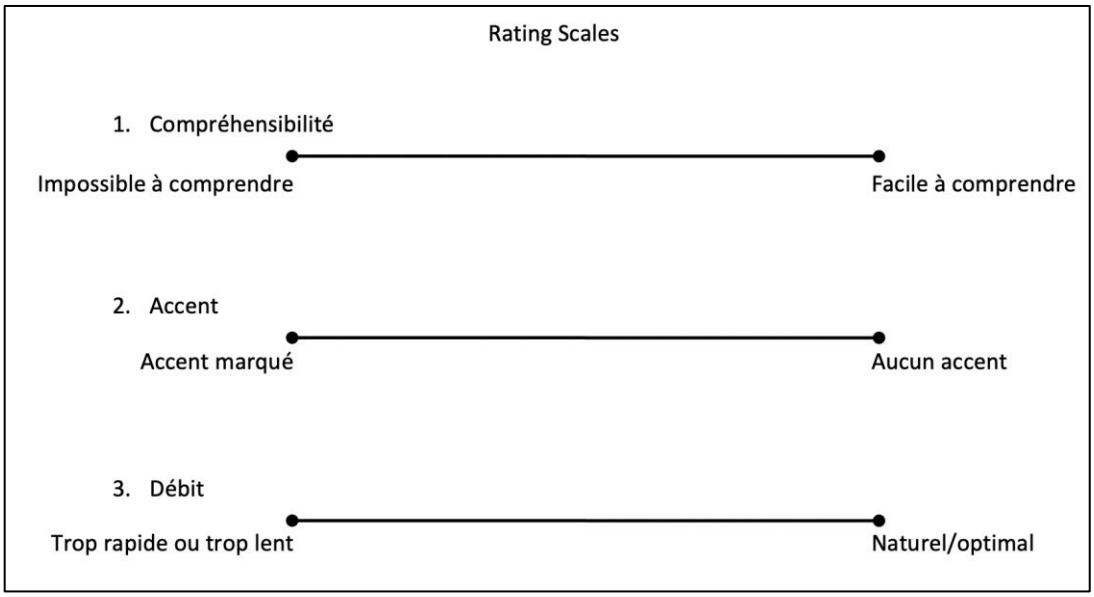

Figure 2. Rating scales used for speech assessments.

External Assessments

External assessments of the students' performances were carried out by 10 native-speaking French female raters (M_{age} = 27.5 years, *SD* = 7.1). The raters were Year 3 or 4 students in a baccalaureate program in teaching French as a second language and were thus representative of the instructors who might be teaching and assessing L2 French students similar to those included as participants here. The raters, who reported knowledge of English, in addition to Spanish (6),

Chinese (1), Ewondo (1), and Italian (1), had taken university-level coursework where phonology or phonetics were covered but had not received formal training in L2 pronunciation teaching. In response to open-ended questions in a background questionnaire, they had reported some teaching experience (M = 1.8 years, SD = 3.2); they also estimated their daily exposure to foreign-accented French at a mean of 44% (SD = 21.7), using a Likert-type rating scale ranging from 0% (none) to 100% (extensive).

The raters participated in individual rating sessions. After the researcher explained the target constructs (with examples), the raters practiced assigning speech ratings using six practice audios, which were the same audios used for benchmarking in the treatment group (cut down to 45 seconds). Besides receiving study-specific explanations (definitions plus examples) and engaging in a brief rating practice, the raters did not receive any additional training or calibration practice, which is consistent with the idea that accentedness, comprehensibility, and fluency reflect listeners' intuitive, holistic judgements of L2 speech (Munro & Derwing, 2015). The raters then proceeded to evaluate Presentation 1 and Presentation 2 audios (45-second samples, the same as those self-assessed by the students), organized in two sets, with all Presentation 1 audios from the treatment and comparison groups included in one set and all Presentation 2 audios from both groups in the other set. The audios were presented in a unique random order per rater, and the order of sets was counterbalanced, such that five raters evaluated Presentation 1 before Presentation 2, and five raters evaluated the sets in the opposite order. The audios were played to the raters through a high-quality headset on a computer, but the raters assigned their ratings in paper booklets, through the same format as the students (see Figure 2). The raters could listen to each audio only once before providing their ratings.

Data Analysis

The students' and the external raters' evaluations were expressed as numerical values (out of 100) by measuring the distance (in millimeters) between the left anchor point on each scale and the rating. Because reliability (Cronbach's alpha) values for external raters' evaluations of accentedness ($\alpha = .83$), comprehensibility ($\alpha = .84$), and fluency ($\alpha = .83$) exceeded the threshold of .70−.80 (Larson-Hall, 2016), the raters' scores were averaged to derive a single mean score for each rated dimension, separately per student and presentation. To obtain a measure of self-assessment accuracy, an alignment score was computed by subtracting the raters' mean score from each student's self-assessed score (as in Trofimovich et al., 2016), separately at each self-assessment episode (three self-assessments for Presentation 1 and one self-assessment for Presentation 2; see Figure 1). Positive alignment scores indicated that the students overestimated their performance, while negative scores suggested that the students underestimated their performance, relative to the external raters' judgements. The closer the alignment scores approached 0, the more aligned the students' and the external raters' assessments were.

Because the task characteristics of Presentations 1 and 2 were different (in that Presentation 1 was largely unscripted and unrehearsed and elicited mostly spontaneous speech whereas Presentation 2 included an extended preparation component and likely involved structured, prepared speech), the alignment scores for Presentations 1 and 2 were never compared directly. Instead, to answer the research question, which asked if the students' self-assessments became aligned with the external raters' evaluations as a function of benchmarking and peer-assessment, we conducted only between-group comparisons. More specifically, we compared the treatment and the comparison groups' alignment scores at each assessment episode: for Presentation 1 in Week 3 (prior to interventions in the treatment group), for Presentation 1 in Week 6 (after benchmarking in the treatment group), and for Presentations 1

and 2 in Week 12 (after peer-assessment in the treatment group). Our assumption was that the effects of benchmarking and peer-assessments would emerge gradually, as shown through numerically smaller (more calibrated) alignment scores in the treatment than in the comparison group. We expected no between-group differences in alignment scores in Week 3 (before interventions) but anticipated possible effects to emerge in Week 6 (post benchmarking) and in Week 12 (post peer-assessment). Because the alignment scores were non-normally distributed and the sample size was small, non-parametric tests were conducted (Larson-Hall, 2016), with all comparisons adjusted using a Bonferroni correction. We used effect sizes to interpret the results of statistical comparisons, following Plonsky and Oswald's (2014) guidelines, because *p* values are particularly sensitive to sample size and because statistical significance does not always reflect meaningful differences.

Results

Accentedness

As illustrated in Figure 3, regardless of the group, the students generally tended to overestimate their accentedness, relative to the external raters' assessments (by a mean of 10–30 points on a 100-point scale), with all mean alignment scores having positive values.[1]

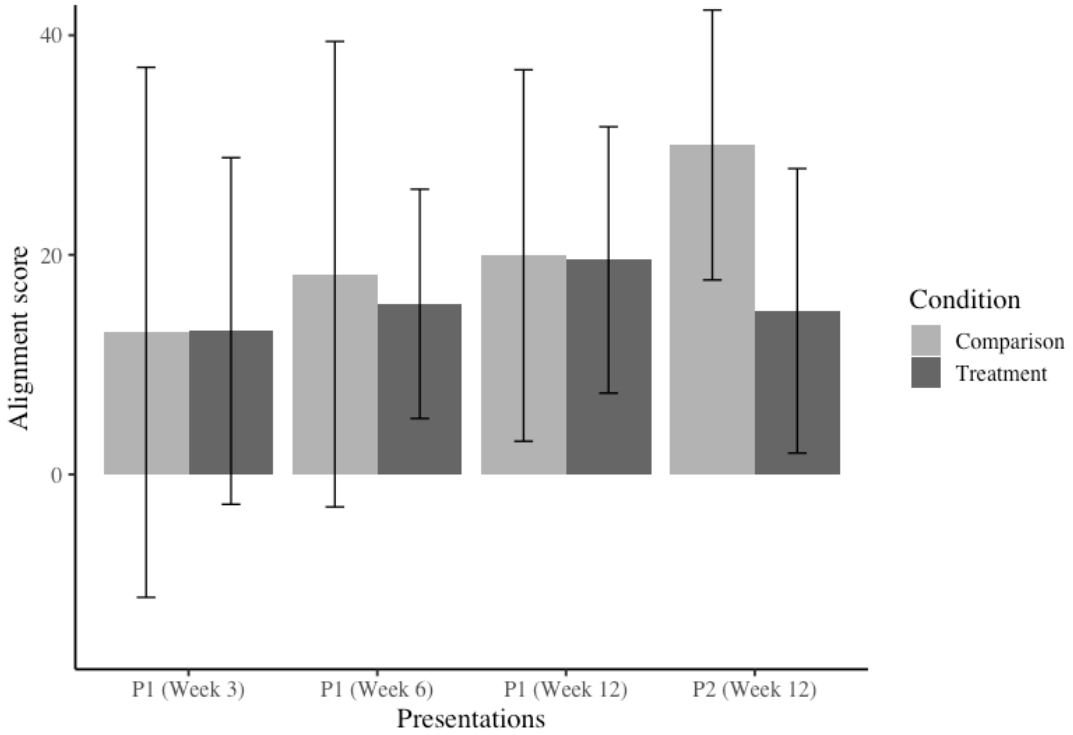

Figure 3. Mean alignment scores for accentedness. P1 = Presentation 1; P2 = Presentation 2. Error bars represent 95% CI.

As summarized in Table 2, Mann-Whitney *U* tests revealed no statistically significant between-group differences, meaning that the students in both groups demonstrated a comparable tendency to overestimate their accentedness. However, the effect size (r = .29) of the comparison for Presentation 2 (Week 12) was weak to medium in strength (Plonsky & Oswald, 2014), implying that the students in the treatment group (who overestimated their accentedness by about 15 points) appeared to show more alignment with the external raters' evaluations than the students in the comparison group (who overestimated their accent by about 30 points; see Figure 3).

Table 2 *Descriptive Statistics and Between-Group Comparisons for Alignment Scores in Accentedness*

	Comparison		Treatment		Between-group comparison			
Rating episode	*M*	*SD*	*M*	*SD*	*U*	*z*	*p*	*r*
Presentation 1 (Week 3)	12.94	33.74	13.07	28.51	78.50	0.19	.849	.04
Presentation 1 (Week 6)	18.24	29.63	15.53	18.85	73.00	–0.11	.935	.02
Presentation 1 (Week 12)	19.94	23.65	19.53	21.91	78.50	0.20	.849	.04
Presentation 2 (Week 12)	30.00	17.19	14.90	23.40	49.00	–1.44	.160	.29

Comprehensibility

As shown in Figure 4, the students again tended to overestimate their comprehensibility, relative to the external raters' assessments, although the range was smaller (by a mean of 3–20 points on a 100-point scale). In some instances, both the comparison group's students (for Presentation 1 in Week 6) and the treatment group's students (for Presentation 2 in Week 12) in fact appeared aligned in their self-assessment with the ratings by the external raters, because their average alignment scores were close to zero.

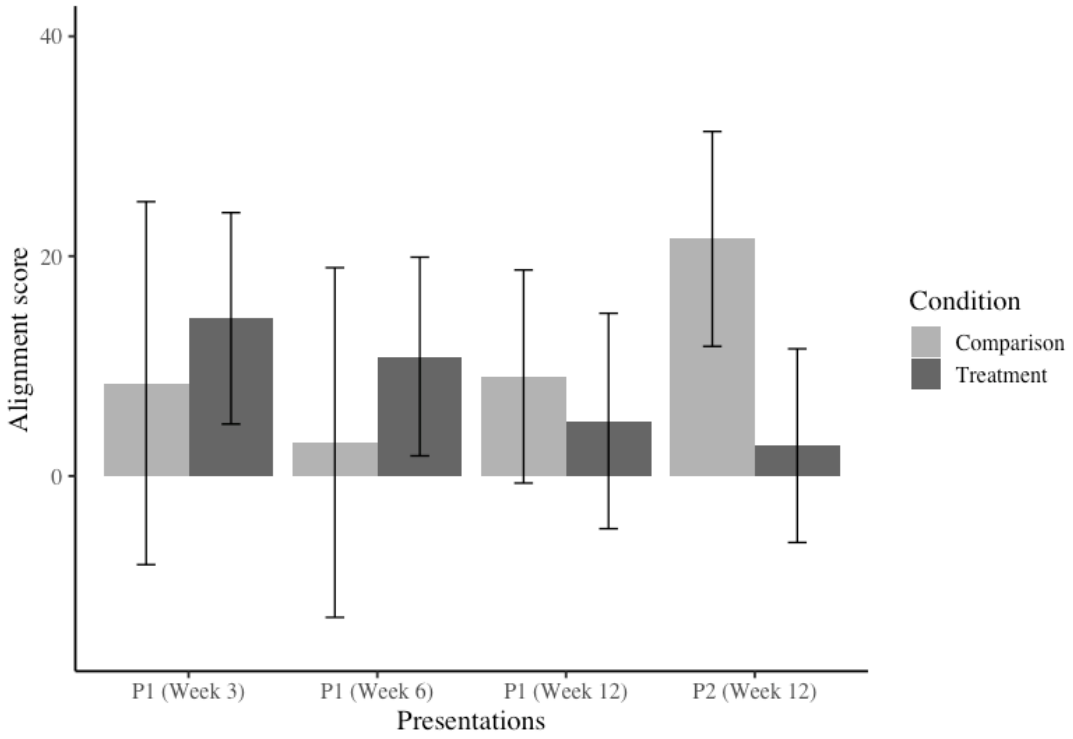

Figure 4. Mean alignment scores for comprehensibility. P1 = Presentation 1; P2 = Presentation 2. Error bars represent 95% CI.

As summarized in Table 3, Mann-Whitney *U* tests revealed no statistically significant differences between the groups for Presentation 1 at any self-assessment episode (Weeks 3, 6, or 12). However, there was a significant between-group difference for Presentation 2 (Week 12), with a medium-to-large effect size (Plonsky & Oswald, 2014), such that the treatment group's students showed a narrower gap between their self-assessments and the external raters' judgments, compared to the comparison group's students. Again, as seen in Figure 4, the treatment group was generally aligned with the external raters' evaluations for Presentation 2.

Table 3 *Descriptive Statistics and Between-Group Comparisons for Alignment Scores in Comprehensibility*

	Comparison		Treatment		Between-group comparison			
Rating episode	*M*	*SD*	*M*	*SD*	*U*	*z*	*p*	*r*
Presentation 1 (Week 3)	8.45	23.07	14.34	17.36	82.00	0.39	.723	.08
Presentation 1 (Week 6)	3.05	22.23	10.87	16.31	94.00	1.05	.311	.21
Presentation 1 (Week 12)	9.05	13.55	5.01	17.69	66.00	–0.50	.643	.01
Presentation 2 (Week 12)	21.57	13.65	2.77	15.91	30.00	–2.50	.012	.50

Fluency

As illustrated in Figure 5, the students for the most part were roughly aligned with the external raters in self-assessments of their fluency for both Presentations 1 and 2, as indicated by the 95% confidence intervals crossing the 0 value in most cases. Although the students tended to overestimate their performance (by a mean of 3–20 points on a 100-point scale), this tendency appeared less pronounced for fluency than for accentedness or comprehensibility (cf. Figures 3–5).

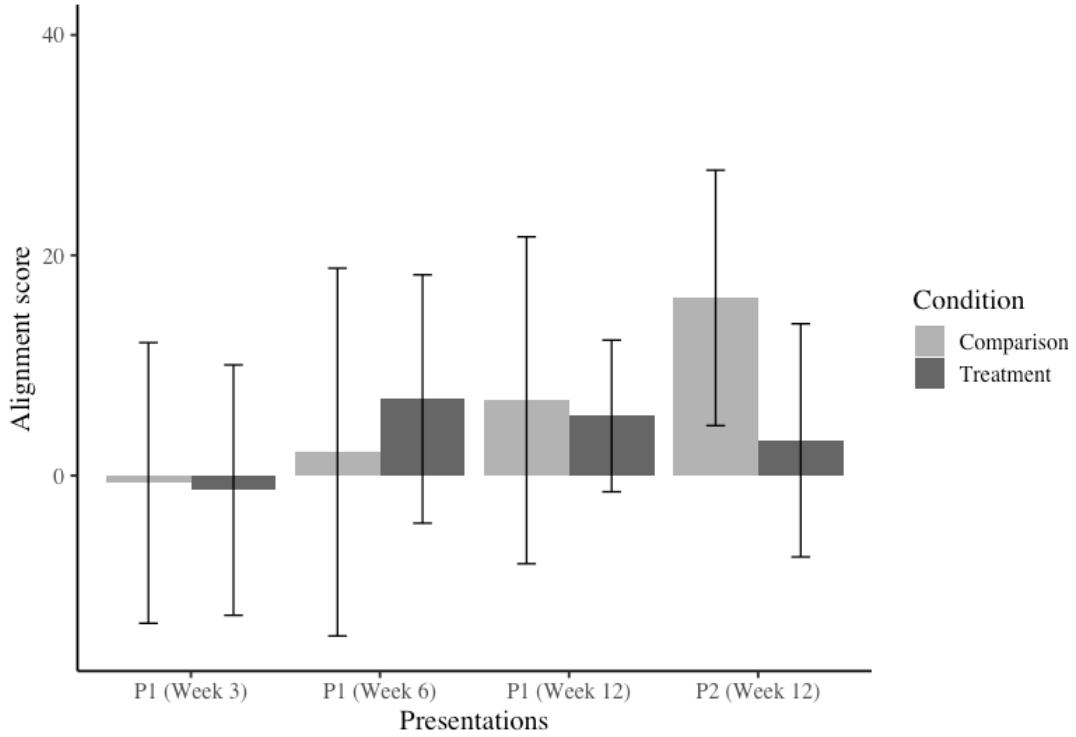

Figure 5. Mean alignment scores for fluency. P1 = Presentation 1; P2 = Presentation 2. Error bars represent 95% CI.

As shown in Table 4, Mann-Whitney *U* tests revealed no between-group differences either for Presentation 1 or Presentation 2. However, as with accentedness, the effect size (*r* = .32) of the comparison for Presentation 2 (Week 12) was weak-to-medium in strength (Plonsky & Oswald, 2014), suggesting a trend for the treatment group (who overestimated their fluency by about 3 points) to show closer alignment with the external raters' evaluations than for the comparison group (who overestimated their fluency by about 16 points; see Figure 5).

Table 4 *Descriptive Statistics and Between-Group Comparisons for Alignment Scores in Fluency*

	Comparison		Treatment		Between-group comparison			
Rating episode	*M*	*SD*	*M*	*SD*	*U*	*z*	*p*	*r*
Presentation 1 (Week 3)	–0.67	17.81	–1.32	20.53	74.00	–0.06	.978	.01
Presentation 1 (Week 6)	2.13	23.33	6.94	20.74	86.00	0.61	.567	.12
Presentation 1 (Week 12)	6.83	20.74	5.41	12.44	75.50	0.03	1.000	.01
Presentation 2 (Week 12)	16.13	12.44	3.19	19.11	46.00	–1.61	.115	.32

Discussion

This study examined the role of benchmarking and peer-assessment activities in L2 French learners' self-assessment of the global dimensions of accentedness, comprehensibility, and fluency in a 15-week L2 French course. Results revealed a difference for the two tasks targeted in this study (Presentation 1 vs. Presentation 2). For Presentation 1 (an oral presentation recorded by the students in Week 3 and self-assessed on three occasions), there were few differences between the treatment and the comparison groups regarding how closely their self-assessments aligned with the ratings provided by the external raters. However, for Presentation 2 (an oral presentation delivered at the end of the course), the treatment group showed greater alignment in their self-assessments of comprehensibility, compared to the comparison group. Because only the treatment group had engaged in benchmarking and peer-assessment activities before Presentation 2, a cautious conclusion emerging from these findings is that focused self- and other-assessment activities (benchmarking, peer-assessment) might have impacted the degree to which the students in the treatment group were able to align their self-assessments of comprehensibility with the ratings provided by the external raters.

Pedagogical Interventions and Self-Assessment Accuracy

Our primary interest was to determine if benchmarking and peer-assessment activities implemented in an L2 French classroom could enhance alignment between learners' self-assessment and other-assessment. In Week 3 (before any pedagogical interventions in the treatment group), both classes were roughly comparable in their self-assessments, relative to external listeners' judgments, implying that any differences emerging in the treatment group beyond this assessment episode could be attributed to the interventions. We saw no clear evidence that the benchmarking activity alone impacted the learners' self-assessments, as the classes performed similarly in the assessment episode that followed the benchmarking activity (in Week 6). Benchmarking activities are considered beneficial because they help learners internalize assessment standards in reference to a range of performance levels in a target skill (Hendry et al., 2009). Our findings of minimal immediate effects of benchmarking on self-assessment depart from the results of Babaii et al. (2016), where discussing and practicing agreed-upon rating criteria with L2 English learners led to improvement in their self-assessment of speaking. However, our findings agree with the results of Peirce et al. (1993), where the use of benchmarks had little impact on L2 French learners' self-evaluation of speaking. The value of benchmarking likely resides in the intensity and the extent of learner engagement in developing, internalizing, and applying the agreed-upon rating criteria. In Babaii et al. (2016), the benchmarking activities spanned several lengthy sessions, where the learners not only developed and discussed their personal rating criteria, based on the evaluations of their own speech, but were also familiarized with the criteria developed by teachers, with researchers modeling their use by evaluating two sample performances in a think-aloud procedure. In contrast, in Peirce et al. (1993), the benchmarking criteria were relatively abstract (e.g., comparing one's performance

relative to a general francophone speaker), and they were shared with learners, without much discussion. In the present study, although a 30-minute benchmarking activity would have allowed the learners to become familiar with various course-specific performance levels, the activity was likely far too brief to enable the learners to internalize concrete assessment benchmarks or to gain sufficient practice in applying these benchmarks consistently.

In this study, evidence of pedagogical interventions having a positive impact on students' self-assessments emerged in Week 12, in student evaluations of Presentation 2, where the treatment group showed greater alignment in their self-assessments of comprehensibility, compared to the comparison group (see Figure 4). This between-group difference most likely arose due to the extensive peer-assessment practice that the students engaged in during Weeks 7–10 while observing and evaluating their classmates' presentations, with every student evaluating everyone else's presentation. Previous work has consistently showed positive effects of peer-assessment on learners' self-assessments of various language skills (e.g., Chen, 2008; Patri, 2002), generally through "an iterative process when [learners] subconsciously compare peers with themselves" (Chen, 2008, p. 244). Our findings further support this conclusion, suggesting that exposure to fellow classmates' presentations, with an added focus on peer-assessment, provided the treatment group with a broad range of pronunciation performance exemplars that they could use as benchmarks in evaluating their own speech. A similar exposure in the comparison group, but without peer-assessment, was insufficient to enable those students to align their self-assessments with those of the external raters, highlighting an important role of peer-assessment in directing students' attention to similarities and differences between their own and their classmates' speech as a way of developing enhanced awareness of their pronunciation. And because the positive effect of peer-assessment was detected only for Presentation 2, not

Presentation 1, it is likely that the benefits of peer-assessment are greatest when peer- and self-assessments are carried out on the same speaking task, which maximizes the transfer of task-specific evaluation criteria as learners compare their own versus their peers' performances.

Learners' Self-Assessments and Global Dimensions of L2 Speech

One incidental finding of this study was that all students (regardless of group assignment) appeared to differ in their self-assessment across the three global dimensions, such that the students' self-ratings of fluency were most similar, whereas their self-ratings of accentedness were least similar, to the evaluations by the external raters. To the best of our knowledge, this finding is novel because no prior self-assessment research has focused on several global L2 speech measures simultaneously. The students' relatively greater ease with self-assessment of comprehensibility and fluency could primarily be attributed to the learning context. The target course was designed to facilitate the development of intelligible and comprehensible speech, with an explicit focus on connected speech processes (e.g., liaison, enchainement), shown to be relevant to listener perception of comprehensibility (Trofimovich et al., 2017). In addition, the definition of fluency (e.g., speech characterized by few pauses and hesitation and an optimal speaking rate) as used here was likely more transparent and therefore understandable to the students, in the sense that the criteria for considering a performance as fluent versus dysfluent were easier to internalize than those for the ratings of comprehensibility and particularly accentedness (e.g., speech characterized by multiple influences). In contrast, accentedness is likely an especially opaque dimension for learners, who often have little understanding of which specific linguistic features make their speech more versus less accented (Derwing, 2003). Moreover, speakers generally tend to prefer their own accent (e.g., Kinzler et al., 2007), evaluating their own accent more favourably than the accent of others (Mitterer et al., 2020),

which is consistent with the tendency for students in this study to overestimate their accentedness. At least for the intermediate-level L2 French learners here, our interim conclusion (which awaits confirmation in future research) is that fluency and comprehensibility were easier to self-assess than accentedness. If perceptive self-assessment facilitates learning, then learner difficulty with self-assessment of accentedness is particularly telling, given that many L2 speakers aspire to attain nativelike, non-accented L2 performance (Derwing, 2003; Inceoglu, 2019).

Limitations and Future Directions

The present findings must remain preliminary until investigated in other teaching contexts and languages and with larger cohorts of learners. Because the students in the treatment group received both benchmarking and peer-assessment activities, we cannot unambiguously attribute the obtained effects in the treatment group to a single treatment activity. For instance, it is possible that the benchmarking activity had a delayed and/or cumulative effect, interacting with repeated peer-assessments, in the sense that the effectiveness of peer-assessment at the end of the course was amplified by the students engaging in benchmarking earlier on. Thus, future work should disentangle individual contributions of specific assessment-focused interventions, which was not possible here. Furthermore, the practical considerations of conducting a project in an existing course prevented us from fully matching the target speaking tasks that were self-evaluated by the students with the performances included in the assessment-focused activities. For example, the types of speaking performances evaluated through the benchmarking and peer-assessment activities were similar to those recorded by the students for Presentation 2; however, the content, context, and requirements for Presentation 1 were different, in that this presentation's topic was chosen by the teacher, the presentation was recorded in a lab, and it was

not meant to be heard by classmates. Moreover, practical considerations also prevented us from targeting tasks that would be comparable in their demands on L2 learners. Presentation 1, which was largely unscripted and unrehearsed, elicited mostly spontaneous speech from the students. In contrast, Presentation 2, which included an extended preparation component involving in-class and at-home practice, elicited structured, prepared, and likely rehearsed speech delivered in front of a live audience. Thus, in future research, task type should be given specific consideration, to understand how benchmarking and peer-assessment activities benefit L2 learners' self-assessments across similar and different speaking tasks. Similarly, the course instructor in both the treatment and control groups was one of the researchers, which may have inadvertently introduced an unwanted researcher bias toward her teaching approach or treatment of individual students. Future studies, therefore, may need to ensure that the target instruction is delivered by teachers who are not also members of research teams (e.g., Meritan & Mroz, 2019). Finally, a close examination of peer-assessment activities, where learners engage in think-alouds, group discussions, or interviews, would help researchers and teachers to understand which aspects of peer-assessment are particularly beneficial for facilitating self-assessment accuracy.

Pedagogical Implications

The current findings offer several practical suggestions for the use of assessment-focused activities. Because peer-assessments of live class presentations appeared beneficial for L2 learners' self-assessment in this study, which is consistent with a growing body of literature on self-assessment (Li & Zhang, 2021), teachers might wish to encourage their students to self-reflect on their own pronunciation performance and compare it with the performance by peers during various classroom activities, including interactive, dialogic tasks or whole-class or small-group presentations. Alternatively, self-assessment or self-reflection activities can be assigned

for practice outside class, with or without the support of relevant computer-assisted adaptive technologies (for examples, see Li & Zhang, 2021). Exchanging feedback between students might also enhance the value of peer-assessment (Chen, 2008; but see Patri, 2002). Although some teachers might feel reluctant to include student self-assessment in their teaching practice, feeling overwhelmed by the need to cover course content within rigid time constraints, self- and peer-assessment activities might help learners (and their teachers) reap multiple benefits in the future. One such benefit, besides students' increased awareness of their own pronunciation strengths and weaknesses, is enhanced learner agency, which might enable students to develop into self-directed, self-motivated learners who seek additional opportunities for language learning and use outside classroom instruction (e.g., Oscarson, 1989). Students with more developed self-assessment skills might also avoid feeling overly frustrated (e.g., Freund & Kasten, 2011), unable to understand the reactions that they might overtly or covertly receive from their interlocutors who may have difficulty understanding students or may find students' accents particularly prominent.

Conclusion

This study's results revealed that peer-assessment activities were generally effective at narrowing the gap between L2 French learners' self-ratings of comprehensibility and the evaluations of comprehensibility by the external raters. Additionally, with its focus on three global dimensions of L2 pronunciation, this study showed that L2 French learners demonstrate various degrees of difficulty with these dimensions, where self-assessments of accentedness might be particularly problematic and self-ratings of fluency may appear least challenging. Although these findings are preliminary until replicated in follow-up work, they underscore the promise of pedagogical interventions as a catalyst in allowing L2 learners to enhance awareness

of their own pronunciation and ultimately to play an active role in their own language learning.

Note

1 An anonymous reviewer raised the concern that individual differences among the external raters may have impacted their assessments of L2 accentedness, such that the raters who displayed different degrees of familiarity with accented L2 French may have been more or less lenient in assessing L2 accentedness (Harding, 2011). To address this concern, prior to reporting and interpreting our findings, we computed Spearman correlations between the raters' self-assessed familiarity with accented L2 French and the accentedness ratings they assigned to the students in both Presentation 1 (r_s = .543, p = .105) and Presentation 2 (r_s = .228, p = .526). Although these (non-significant) associations were weak-to-moderate in strength, a visual inspection of the data did not reveal clearly interpretable patterns, so no further analyses focusing on individual differences in the raters' background profiles were pursued. Such analyses would additionally fall far beyond the remit of this study.

Acknowledgements

We are grateful to Cynthia Lapierre and Kaitlin Vida Kelly for their help with data collection and data analysis. An earlier version of this work was presented at the 6th International Conference on English Pronunciation: Issues & Practices (EPIP 2019). This research was supported by grants from the Social Sciences and Humanities Research Council of Canada (SSHRC) to Pavel Trofimovich (435-2016-1406) and Sara Kennedy (430-2016-0441).

References

ACTFL. (2012). *ACTFL Proficiency Guidelines 2012* [Electronic version]. Retrieved from https://www.actfl.org/sites/default/files/pdfs/public/ACTFLProficiencyGuidelines2012_FINAL.pdf

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing*, *33*, 411–437. https://doi.org/10.1177/0265532215590847

Butler, Y. G., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, *27*, 5–31. https://doi.org/10.1177/0265532209346370

Chen, Y.-M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research*, *12*, 235–262. https://doi.org/10.1177/1362168807086293

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

Denies, K., & Janssen, R. (2016). Country and gender differences in the functioning of CEFR-based can-do statements as a tool for self-assessing English proficiency. *Language Assessment Quarterly*, *13*, 251–276. https://doi.org/10.1080/15434303.2016.1212055

Derwing, T. (2003). What do ESL students say about their accents? *Canadian Modern Language Review*, *59*, 547–567. https://doi.org/10.3138/cmlr.59.4.547

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research.* John Benjamins.

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

Dolosic, H. N., Brantmeier, C., Strube, M., & Hogrebe, M. C. (2016). Living language: Self-assessment, oral production, and domestic immersion. *Foreign Language Annals*, *49*,

302–316. https://doi.org/10.1111/flan.12191

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. https://doi.org/10.1111/j.1529-1006.2004.00018.x

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing*, *24*, 37–64. https://doi.org/10.1177/0265532207071511

Freund, P. A., & Kasten, N. (2011). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*, 296–321. https://doi.org/10.1037/a0026556

Georgiou, G. P. (2019). EFL teachers' cognitions about pronunciation in Cyprus. *Journal of Multilingual and Multicultural Development, 40*, 538–550. https://doi.org/10.1080/01434632.2018.1539090

Hansen Edwards, J. G. (2013). Peer assessment in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 730–750). Wiley-Blackwell.

Harding, L. (2011). *Accent and listening assessment: A validation study of the use of speakers with L2 accents on an academic English listening test*. Peter Lang.

Hendry, G. D., Bromberger, N., & Armstrong, S. (2009). Constructive guidance and feedback for learning: The usefulness of exemplars, marking sheets and different types of feedback in a first year law subject. *Assessment & Evaluation in Higher Education, 36*, 1–11. https://doi.org/10.1080/02602930903128904

Hung, Y. (2019). Bridging assessment and achievement: Repeated practice of self-assessment in college English classes in Taiwan. *Assessment & Evaluation in Higher Education, 44*,

1191–1208. https://doi.org/10.1080/02602938.2019.1584783

Inceoglu, S. (2019). Exploring the effects of instruction on L2 French learner pronunciation, accentedness, comprehensibility, and fluency: An online classroom study. *Journal of Second Language Pronunciation, 5*, 223–246. https://doi.org/10.1075/jslp.18004.inc

Kennedy, S., & Trofimovich, P. (2019). Comprehensibility: A useful tool to explore listener understanding. *Canadian Modern Language Review, 75*, 275–284. https://doi.org/10.3138/cmlr.2019-0280

Kinzler, K. D., Dupoux E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 12577–12580. https://doi.org/10.1073/pnas.0705345104

Kissling, E. M., & O'Donnell, M. E. (2015). Increasing language awareness and self-efficacy of FL students using self-assessment and the ACTFL proficiency guidelines. *Language Awareness, 24*, 283–302. https://doi.org/10.1080/09658416.2015.1099659

Lappin-Fortin, K., & Rye, B. J. (2014). The use of pre-/posttest and self-assessment tools in a French pronunciation course. *Foreign Language Annals, 47*, 300–320. https://doi.org/10.1111/flan.12083

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. Routledge.

Lee, S. K., & Chang, S. (2005). Learner involvement in self-and peer-assessment of task-based oral performance. *Language Research, 41*, 711–735.

Levis, J. M. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation, 6*, 310–328. https://doi.org/10.1075/jslp.20050.lev

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in

language testing and assessment. *Language Testing, 38,* 189–218. https://doi.org/10.1177/0265532220932481

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 2*6, 75–100. https://doi.org/10.1177/0265532208097337

Meritan, C. (2020). *The impact of self-reflection and integrated pronunciation instruction on the intelligibility of Generation Z learners of French: A mixed methods study.* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Meritan, C., & Mroz, A. (2019). Impact of self-reflection and awareness-raising on novice French learners' pronunciation. *Foreign Language Annals*, *52,* 798–821. https://doi.org/10.1111/flan.12429

Mitterer, H., Eger, N. A., & Reinisch, E. (2020). My English sounds better than yours: Second-language learners perceive their own accent as better than that of their peers. *Plos One, 15*(2), e0227643. https://doi.org/10.1371/journal.pone.0227643

Murphy, J. M. (2014). Intelligible, comprehensible, non-native models in ESL/EFL pronunciation teaching. *System, 42,* 258–269. https://doi.org/10.1016/j.system.2013.12.007

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation*, *5,* 294–323. https://doi.org/10.1075/jslp.18016.nag

Oscarson, M. (1989). Self-assessment of language proficiency: Rationale and applications. *Language Testing*, *6,* 1–13. https://doi.org/10.1177/026553228900600103

Patri, M. (2002). The influence of peer feedback on self- and peer-assessment of oral skills. *Language Testing*, *19,* 109–131. https://doi.org/10.1191/0265532202lt224oa

Peirce, B. N., Swain, M., & Hart, D. (1993). Self-assessment, French immersion, and locus of control. *Applied Linguistics*, *14*, 25–42. https://doi.org/10.1093/applin/14.1.25

Plonsky, L., & Oswald, F. L. (2014). How big is "big?" Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *Canadian Modern Language Review, 69,* 324–348. https://doi.org/10.3138/cmlr.1748

Strachan, L., Kennedy, S., & Trofimovich, P. (2019). Second language speakers' awareness of their own comprehensibility: Examining task repetition and self-assessment. *Journal of Second Language Pronunciation, 5*, 347–373. https://doi.org/10.1075/jslp.18008.str

Sturm, J. L. (2013). Attention, awareness, and accents in L2 French. *Language Awareness, 22*, 146–160. https://doi.org/10.1080/09658416.2012.670243

Sturm, J. L. (2019). Current approaches to pronunciation instruction: A longitudinal case study in French. *Foreign Language Annals, 52*, 32–44. https://doi.org/10.1111/flan.12376

Suzuki, S., & Kormos, J. (2020). Linguistic dimensions of comprehensibility and perceived fluency: An investigation of complexity, accuracy, and fluency in second language argumentative speech. *Studies in Second Language Acquisition, 42*, 143–167. https://doi.org/10.1017/S0272263119000421

Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing, 32*, 63–81. https://doi.org/10.1177/0265532214541885

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*, 249–276. https://doi.org/10.3102/00346543068003249

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility.

Bilingualism: Language and Cognition, 15, 905–916. https://doi.org/10.1017/S1366728912000168

Trofimovich, P., Isaacs, T., Kennedy, S., Saito, K., & Crowther, D. (2016). Flawed self-assessment: Investigating self- and other-perception of second language speech. *Bilingualism: Language and Cognition, 19*, 122–140. http://doi.org/10.1017/S1366728914000832

Trofimovich, P., Kennedy, S., & Blanchet, J. (2017). Development of second language French oral skills in an instructed setting: A focus on speech ratings. *Canadian Journal of Applied Linguistics*, *20*, 32–50. https://journals.lib.unb.ca/index.php/CJAL/article/view/24573