

Dunning-Kruger effect in second language speech learning: How does self perception align with other perception over time?



Kazuya Saito^{a,*}, Pavel Trofimovich^b, Mariko Abe^c, Yo In'nami^c

^a University College London, London, UK

^b Concordia University, Montreal, Canada

^c Chuo University, Tokyo, Japan

1. Introduction

The Dunning–Kruger effect—a phenomenon documented across various skills, such as people's capacity to solve math problems or to detect grammar errors (Dunning, 2011)—describes the tendency for unskilled performers to overestimate their ability relative to external assessments. Skilled performers, who are often equally inaccurate, instead tend to underestimate their performance. This mismatch between self- and other-assessments, which can preclude people from making skill-appropriate decisions, such as whether to pursue a specific career (e.g., Dunning, Heath, & Suls, 2004), is attributed (at least for poor performers) to people suffering from lack of meta-knowledge. That is, a low level of skill enables poor performers to make mistakes but also prevents them from recognizing that more skilled individuals may act or perform differently (Schlösser, Dunning, Johnson, & Kruger, 2013). Our goal in this study was to investigate the Dunning–Kruger effect with respect to how second language (L2) speakers' assess comprehensibility (ease of understanding) of their L2 speech.

2. Background literature

Examining the Dunning–Kruger effect in relation to L2 speakers' assessments of their speech is important for both conceptual and pedagogical reasons. From a conceptual standpoint, several theoretical accounts of L2 development posit that the linguistic system develops when L2 speakers notice and subsequently minimize the gap between the target linguistic system and the speakers' own conception of it (e.g., see Ellis, 1997, for a computational model, and Schmidt, 2001, for a noticing framework). Therefore, to facilitate L2 acquisition, it would be important to understand whether and at which point in the learning process L2 speakers can adequately assess their performance. And from a pedagogical perspective, L2 speakers' self-assessment has been used to complement various types of evaluations in many language classrooms (Babaii, Taghaddomi, & Pashmforoosh, 2016; Kissling & O'Donnell, 2015; Lappin-Fortin & Rye, 2014). Thus, to test the robustness of classroom assessment practices, it is crucial to probe whether and at

which point in the learning process self-assessments may accurately reflect at least some aspects of L2 speakers' performance (Suzuki, 2015).

Prior research on self-assessment of L2 speakers' speaking skills has largely focused on the links between assessment by instructors and speakers' self-evaluations (typically for university students). This research has often revealed only weak relationships between L2 speakers' assessment of their performance in a task and their instructor's evaluation of the same task (Babaii et al., 2016) or between L2 speakers' assessment of specific features of their speech (e.g., vowels, intonation patterns) and their teacher's assessment of the same features (Lappin-Fortin & Rye, 2014). However, after L2 speakers receive training in self-assessment, correlations between self- and other-assessments seem to improve. For instance, when speakers discuss their performance or become familiar with rating criteria, their self-evaluations become more specific, detailed, and constructive and tend to converge with instructors' evaluations (Babaii et al., 2016; Kissling & O'Donnell, 2015).

Although L2 speakers' inaccurate self-assessments may be remedied, it is unknown whether self-assessments become more calibrated with external evaluations *over time*, without targeted training in self-assessment, and whether there is individual variation in how easily L2 speakers' self-assessments align with external evaluations. At a basic level, it is also unclear whether L2 speakers' inaccurate self-assessments (those consistent with the Dunning–Kruger effect) extend beyond their evaluations of specific task performances or particular features of speech to include global dimensions, such as comprehensibility, which refers to how easy it is for a listener to understand a speaker (Munro & Derwing, 1995).

As a global dimension of L2 speech, comprehensibility is a practical and useful measure of L2 speakers' performance, complementing such specific metrics as speakers' accurate and fluent use of phonological, lexical, and grammatical aspects of language (see Nagle, Trofimovich, & Bergeron, 2019; Crowther, Trofimovich, Saito, & Isaacs, 2015; Saito, Trofimovich, & Isaacs, 2017; Saito & Plonsky, 2019). For example, listeners typically demonstrate high consistency in their comprehensibility judgments (Munro, 2018; see Saito & Plonsky, 2019 for meta-

* Corresponding author at: University College London, Institute of Education, 20 Bedford Way, WC1H 0AL, UK.
E-mail address: k.saito@ucl.ac.uk (K. Saito).

analysis). By contrast, listeners' actual understanding of L2 speech—captured through such intelligibility measures as transcription accuracy or narrative retell—is often highly dependent on the measurement task (Kang, Thomson, & Moran, 2018; Kennedy, 2009) and on the nature of the speech sample (i.e., speech samples should be unique to avoid greater intelligibility for repeated content). Comprehensibility is also a popular measure of L2 speech, featured in a variety of high- and low-stakes assessment instruments (e.g., Derwing & Munro, 2015). Furthermore, apart from being practical, reliable, and popular, comprehensibility captures valuable information about listeners' understanding of and reactions to L2 speech. For instance, comprehensibility ratings reflect the time that listeners require to process L2 speech (Ludwig & Mora, 2017; Munro & Derwing, 1995) and listeners' negative, emotional reactions towards L2 speakers if processing effort is high (Dragojevic & Giles, 2016; Lev-Ari & Keysar, 2010). Thus, comprehensibility appears to be an important global dimension of L2 speech to be investigated in relation to L2 speakers' self-assessment, on the assumption that an accurate self-assessment of comprehensibility would enable L2 speakers to estimate their interlocutors' reactions to their speaking performance.

In an exploratory study, Trofimovich, Isaacs, Kennedy, Saito, & Crowther, 2016 recently focused on documenting Dunning–Kruger effects in L2 speakers' self-assessments of comprehensibility. In this cross-sectional dataset, 134 L2 students at an English-medium university in Canada completed a picture narrative task and self-rated their comprehensibility in this task (degree to which they believed their L2 speech was easy to understand for listeners). The students' performance was then evaluated by native English-speaking listeners for the same construct. Comparisons of self- versus other-assessments revealed findings consistent with the Dunning–Kruger effect: L2 speakers at lower comprehensibility levels overestimated their performance while speakers at upper levels underestimated it.

Although these results suggest that L2 comprehensibility, like other constructs in cognitive and social psychology (e.g., Dunning, 2011), is subject to faulty self-assessment, especially at low skill levels, the generalizability of these results to other contexts and other speakers is limited. In particular, the occurrence of the Dunning–Kruger effect with respect to speakers' L2 oral skills naturally raises the questions about the generalizability of these findings to other contexts (e.g., beyond Canada, where English is spoken outside the language classroom), other teaching situations (e.g., classroom instruction), and other timeframes (e.g., longitudinal rather than cross-sectional). With these questions in mind, the objectives of this study were twofold. First, we aimed to replicate the findings of Trofimovich et al., 2016 in a different L2 learner context, namely, with 106 Japanese secondary students in an English as a foreign language (EFL) setting without the opportunity to use L2 English on a daily basis. Second, unlike in the initial study, which involved a single measurement of L2 speakers' self-assessment (Trofimovich et al., 2016), we examined the Dunning–Kruger effect from a longitudinal perspective, investigating whether and to what extent L2 speakers could align their self-perception with the assessment of external listeners during one academic term (three months) of EFL instruction, as a function of speakers' individual difference profiles.

The goal of identifying potential predictor variables—by focusing on L2 speakers' individual difference profiles—was particularly important for us, in the sense that it is crucial not only to document Dunning–Kruger effects in L2 speech learning but also to attempt to explain why some L2 speakers do while others do not succeed in accurate self-assessments. One cluster of individual difference variables that can be useful in explaining L2 speakers' self-assessment behaviors pertains to different types of socio-psychological orientations among L2 speakers. With respect to motivation (Dörnyei, 2005), for example, there is much empirical evidence that L2 speakers' willingness to engage in L2 practice is more strongly associated with their self-image as an ideal future L2 user (Ideal L2 Self) than with their awareness of the characteristics that they must possess so as to succeed in language

learning and to avoid negative outcomes (Ought L2 Self). In terms of emotion, much work has been devoted to understanding the role of negative emotions in classroom L2 learning (i.e., Foreign Language Anxiety) and to clarifying their negative impact on L2 speakers' achievement (e.g., Horwitz, 2017). More recently, emotion researchers have begun to emphasize the role of positive emotions (i.e., Foreign Language Enjoyment), suggesting that it is positive (rather than negative) emotional states that exert substantial amounts of influence on L2 speakers' behaviors and outcomes of L2 learning (e.g., Dewaele & MacIntyre, 2014; Dewaele, Witney, Saito, & Dewaele, 2018).

3. The present study

Drawing on the data from a larger-scale project surveying the longitudinal development of L2 speech in Japanese EFL classrooms, we analyzed first-year high school students' speaking performance for comprehensibility at the beginning (T1) and end (T2) of one academic term (second semester), from December 2016 to March 2017. The primary goal of the project was to create a longitudinal learner speech corpus, wherein researchers can investigate the complex mechanisms underlying classroom L2 speech learning by linking individual students' developmental patterns to a range of biographical variables.

In an earlier report (Saito, Dewaele, Abe, & In'nami, 2018), we investigated the degree of the students' speaking performance gains as a function of three individual difference factors: (a) students' use of English inside and outside class, (b) their L2 motivation (Dörnyei, 2005), and (c) their L2 emotion, both negative and positive (Dewaele & MacIntyre, 2014). Our findings revealed a great deal of individual variability in students' L2 speech development. Students' improvement in comprehensibility was only marginally associated with how much they had practiced using the target language inside and outside classrooms. Rather, students' gains depended on their individual difference profiles, in the sense that those who made the most of their EFL experience were those who expressed positive feelings towards their language learning (i.e., Foreign Language Enjoyment) and had a clearer image of what they wished to achieve in the future through the use of the target language (i.e., Ideal L2 Self).

Extending the dataset, in the current study, we examined the same students' L2 speech learning from a different angle, investigating the alignment between their self-assessments and the evaluation of their speaking performance by external listeners. To this end, we first focused on the new (previously unpublished) dataset—that is, the students' own assessments of their L2 comprehensibility at T1 and T2. We then examined how the relationship between the students' self-assessments and the assessments of their performance by external listeners changed over the three months of the project (i.e., whether the students calibrated their self-assessments with those of external listeners). Finally, to understand the source of individual variability in the students' calibration processes, we evaluated the extent to which any longitudinal changes in self- versus other-assessments could be associated with the students' experiential and socio-psychological individual profiles. The study was guided by the following two questions:

1. Do L2 students become more aligned in their self-assessments of L2 comprehensibility, relative to the judgments of external listeners, during one academic term of EFL studies?
2. Are there individual differences in L2 students' experiential profiles that can explain why some students might be more successful than others at calibrating their self-assessments with those of others?

4. Method

4.1. Participants

Our original dataset included 122 first-year high school Japanese EFL students, but the data for 16 students were eliminated due to low

quality of recordings ($n = 6$) and failure to complete both T1 and T2 speaking tests ($n = 8$) or to provide T1 self-assessments ($n = 2$). The high school from which the students were recruited was private and considered to be prestigious. According to the school's annual report, nearly all graduating students take entrance exams to continue their studies at university. Based on the results of the background questionnaire, which was administered at the outset of the project, the students included in the final analyses ($n = 106$) differed in age of first exposure to English ($M = 10.1$ years, $SD = 3.0$, $range = 1$ – 14) and total length of instruction ($M = 874.0$ h, $SD = 520.9$, $range = 200$ – 4500), but language experience was restricted for all students to classroom instruction in Japan, as none reported having resided in English-speaking countries except for brief family visits. The participating students' general L2 English proficiency was estimated through general proficiency test scores (The EIKEN Test in Practical English Proficiency) which ranged from Grades Pre–2 to 2. The proficiency scores indicated that students could be considered as Basic through Independent Users (A2–B1) according to the CEFR benchmarks (EIKEN Foundation of Japan, 2017). Before conducting the study, the researchers obtained human subjects approval from the institutional review board at Chuo University, in accordance with the university's human subjects guidelines. Before taking part in research, all participants read an information sheet (written in Japanese) and provided their written consent.

4.2. Individual difference profiles

During the duration of the project (December 2016–March 2017), the students attended seven 50-min EFL classes per week: (a) five lessons entitled *General English* and (b) two lessons entitled *English Production*. These classes were alternatively taught throughout the academic year by the same three Japanese teachers with high-level L2 English proficiency. We had full access to all syllabi and lesson plans for both courses. Additionally, several formal and casual classroom observations were conducted to confirm the content and consistency of the curriculum throughout the project. Based on the evidence, both courses included several comprehension and production tasks, where students were encouraged to use language both accurately and fluently by applying newly learned grammar points and useful expressions through interaction (i.e., focus on form). The overall nature of instruction was highly consistent because the teachers used a standard syllabus and held regular meetings, which ensured a similar EFL classroom experience for all students throughout the project. We provide more detailed information of the English instruction in Appendix A in the Supporting Information online.

To further capture any individual differences in the students' EFL experience during the project (throughout the 3 months between T1 and T2), we developed an EFL Experience Questionnaire (see Appendix B in the Supporting Information online) based on prior EFL studies (e.g., Muñoz, 2014; Saito & Hanzawa, 2016) and asked all students to complete it at the end of the project (T2). The questionnaire asked how much students had used L2 English inside and outside classroom instruction. For language use inside classrooms, they estimated what percentage of time they had spoken L2 English during class. For out-of-class language use, they reported whether (at the time of the project) they practiced English in two contexts: (a) by studying at cram schools¹ and (b) by engaging in conversational activities with native speakers or/and English L2 users. Descriptive statistics for the quantity and quality of students' EFL experience are summarized in Table 1.

The students' motivation and emotion orientations were surveyed via a tailored composite questionnaire at the beginning of the project (T1). Building on the Japanese version of the L2 Motivational Self

Table 1
Summary of participants' EFL experience between T1 and T2.

	<i>M</i>	<i>SD</i>	<i>Range</i>
In-class experience			
Ratio of speaking inside classrooms (%)	55.4	19.3	5–90
Extracurricular practice outside classrooms (no. of hours per week)			
To prepare for classes (hours)	6.0	3.3	0–15.5
At cram schools ^a	Yes (17), No (83)		
Conversation with native and nonnative speakers ^a	Yes (29), No (74)		

Note. Some participants did not report their experience related to schools ($n = 6$) and conversational activities ($n = 3$).

System Questionnaire (Taguchi, Magid, & Papi, 2009) and the Foreign Language Emotion Scale (Dewaele & MacIntyre, 2014), the questionnaire encompassed four items for Ideal L2 Self, four items for Ought-to L2 Self, eight items for Anxiety and 10 items for Enjoyment. The students rated each statement (e.g., “It's cool to know a Foreign language,” “I imagine myself as someone who is able to speak English”) on a 6-point scale (1 = *strongly disagree*, 6 = *strongly agree*). Reliability analyses (computed using the current dataset) revealed high Cronbach's α values for Ideal L2 Self ($\alpha = 0.82$), Ought-to L2 Self ($\alpha = 0.85$), Foreign Language Anxiety ($\alpha = 0.83$), and Foreign Language Enjoyment ($\alpha = 0.86$). These values corresponded to similar reliability indexes reported in the original validation studies (Dewaele & MacIntyre, 2014, for enjoyment, and Taguchi et al., 2009, for motivation). Subsequently, students' individual scores were averaged in accordance with the four broad categories—Ideal L2 Self, Ought-to L2 Self, Anxiety, and Enjoyment. Descriptive statistics for motivation and emotion variables are presented in Table 2.

4.3. Speech materials

One methodological difficulty of conducting longitudinal research, especially with a large number of participants, is to elicit high-quality spontaneous speech from participants within a short timeframe. To address this difficulty, the Telephone Standard Speaking Test (ALC Press Inc., 2017), an automated English test widely used in Japan, was used to derive samples of students' spontaneous speech at T1 and T2. Using landline or cellphone service at their convenience (typically in a quiet location), the students responded to 10 recorded questions in English, with 45 s allotted to each without planning time. To avoid confusion, all instructions were delivered in both Japanese and English. Our analysis focused on responses to Question 7, eliciting information about a past event (e.g., favorite movie, family trip, shopping). Different from the other questions, where participants could speak without much attention to tense marking, Question 7 was considered to be demanding, because it required students to use the simple past tense forms accurately and consistently throughout the task. Simple past is notoriously difficult to acquire for L2 learners (Collins, Trofimovich, White, Cardoso, & Horst, 2009), so Question 7 was assumed to avoid any ceiling effects and reveal variation in L2 comprehensibility, especially because morphosyntactic accuracy is one linguistic dimension relevant to comprehensibility ratings (e.g., Munro & Derwing, 1995).

A different individual topic targeting past events (out of 17 alternatives) was randomly assigned to each student (with approximately six students per topic) at T1 and T2. This approach allowed us to avoid any test–retest effects (in the sense that students never engaged in the same topic twice). However, there was a possibility that different topics could result in different levels of difficulty. To this end, a one-way ANOVA was performed with students' comprehensibility scores (see below) as the dependent variable focusing on topic type (17 alternatives) at T1 and T2, respectively. The results showed that student performance did not significantly differ across topics at T1, $F(16, 89) = 1.28$, $p = .28$, η_p^2 (effect size) = 0.18, or at T2, $F(16, 89) = 1.43$, $p = .13$, $\eta_p^2 = 0.21$,

¹ In Japan, many high school students choose to go to cram schools in order to prepare for high school exams and for future university entrance exams.

Table 2
A summary of students' motivation and emotion orientations.

Construct	Items (<i>k</i>)	<i>M</i>	<i>SD</i>	<i>Range</i>
Ideal L2 Self	4	3.4	1.3	1.0–6.0
Ought-to L2 Self	4	3.0	1.1	1.0–6.0
Anxiety	8	3.5	0.9	1.3–6.0
Enjoyment	10	4.5	0.8	2.3–6.0

suggesting that students' L2 speech performance was unrelated to the types of topics targeted in the assessment task.

4.4. Comprehensibility rating sessions

Shortly after testing, the students' responses to Question 7 were excised from the recordings, trimmed to the initial 30 s, and normalized for loudness. Following prior research (Derwing & Munro, 2015), five native English-speaking listeners (students at a UK university, all trained in applied linguistics) participated in individual sessions to provide external assessments. First, from a trained research assistant, they received a brief explanation of the key objectives of the study (examining L2 comprehensibility development of Japanese EFL students) and were told what characterized L2 comprehensibility in this study. Building on Derwing and Munro's work, the following training script was used: "Comprehensibility refers to how much effort it takes to understand what someone is saying. If you can understand with ease, then a speaker is highly comprehensible. However, if you struggle and must listen very carefully, or in fact cannot understand what is being said at all, then a speaker has low comprehensibility."

The listeners then practiced the rating procedure by using three similar speech samples, which were not included in the main dataset. Once the listeners felt comfortable, they evaluated all recordings (from both testing times), played in a randomized order, rating them for comprehensibility on a 9-point scale (1 = *difficult to understand*, 9 = *easy to understand*) in a self-paced task with one listening allowed per file. Because the five listeners demonstrated substantial agreement (Cronbach's $\alpha = 0.90$), their ratings were averaged per student to derive separate T1 and T2 scores. Comprehensibility scores varied widely across students at T1 ($M = 4.5$, $SD = 1.7$, $range = 1-9$) and T2 ($M = 4.3$, $SD = 1.5$, $range = 1-9$), implying that students represented a range of comprehensibility levels at both testing times.

4.5. Self-assessment sessions

Within 10 days of the T1 and T2 tests, the students were asked to

self-rate their own comprehensibility on a 10-point scale (1 = *difficult to understand*, 10 = *easy to understand*). We had initially planned to use the same 9-point scale as in much of previous work on L2 comprehensibility; however, when we conducted a pilot study followed by an individual interview, most high school students (as language speakers with little experience in L2 assessment) tended to choose the midpoint value (5 out of 9) when it was available. Therefore, we decided to adopt a 10-point scale to encourage students to use the entire scalar range. To help students understand the concept of comprehensibility, they were given the following script in Japanese: "Comprehensibility refers to how much effort your conversational partners need to make to understand what you are saying in English, despite detectable Japanese accents." Students were also asked to evaluate their comprehensibility without explicitly comparing themselves to their other classmates (for the same procedure, see Trofimovich et al., 2016). Although the 10-point scale was given, no students rated themselves as 10 at T1 ($M = 4.2$, $SD = 1.7$, $range = 1-9$) or T2 ($M = 3.7$, $SD = 1.6$, $range = 1-9$).

5. Results

5.1. Relationships between self- and other-assessments

To enable meaningful comparisons, self-assessments (10-point scale) and external assessments (9-point scale) were converted to z scores, which allowed for direct comparisons of self- and other-perception scores. According to one-sample Kolmogorov-Smirnov tests, the students' self-ratings and the listeners' assessments were normally distributed ($p > .05$). To examine relationships between self- and other-assessments, Pearson correlations were computed between the z scores of the students' self-ratings and external listeners' assessments (Fig. 1). Whereas the two measures were not correlated at T1, $r = 0.07$, $p = .43$, a weak-to-medium strength association emerged at T2, $r = 0.31$, $p = .001$ (Plonsky & Oswald, 2014), implying that the two rating sets were aligned more closely at T2 than T1.

Following prior research (Trofimovich et al., 2016), overconfidence scores were calculated by subtracting, for each student at each testing time, the mean external listeners' comprehensibility score from the student's self-assessment, such that positive scores designated self-ratings that overestimated the student's comprehensibility, negative scores described self-ratings that underestimated the student's performance, and scores around zero corresponded to calibrated assessments. As visually presented in Fig. 2, there were also strong negative associations between students' overconfidence scores and external listeners' ratings at T1, $r = -0.68$, $p < .001$, and T2, $r = -0.58$, $p < .001$. At least at

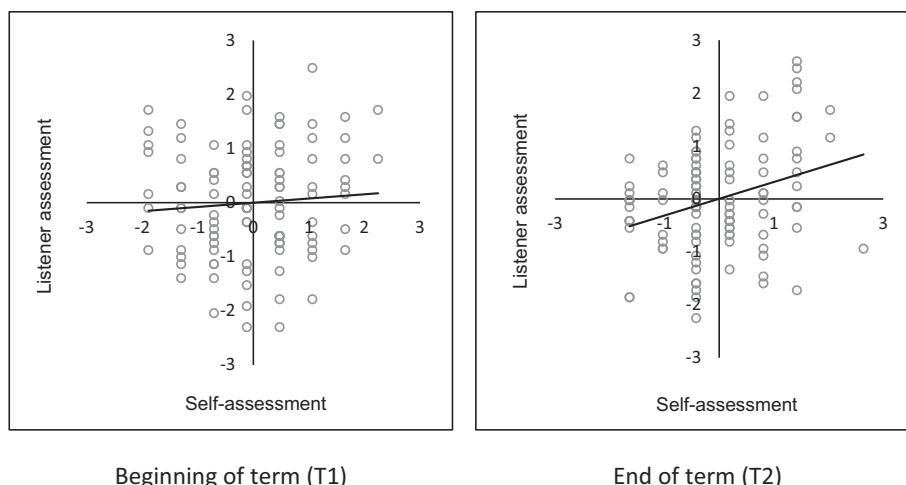


Fig. 1. Relationship between self-assessments (x-axis) and external listeners' assessments (y-axis) at T1 (left) and T2 (right).

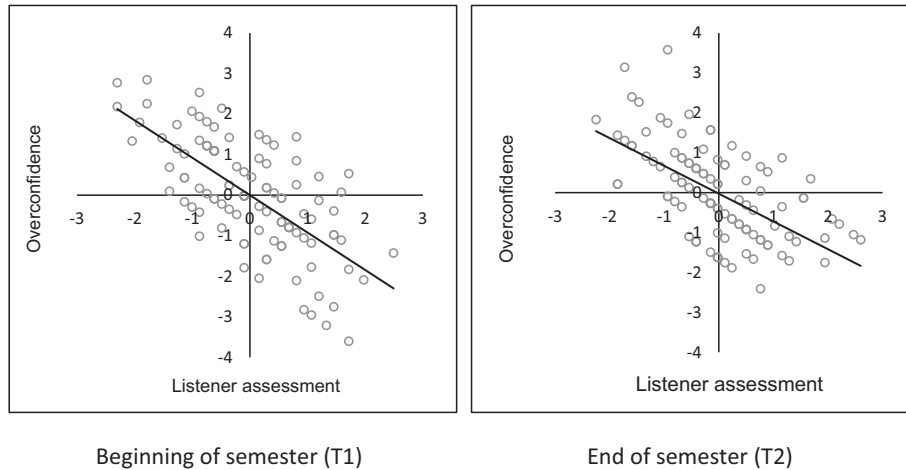


Fig. 2. Relationship between external listeners' assessments (x-axis) and the students' overconfidence scores (y-axis) at T1 (left) and T2 (right).

first glance, these associations could be explained through (a variation of) the regression-to-the-mean effect (Burson, Larrick, & Klayman, 2006; Feld, Sauermann, & de Grip, 2017), in the sense that lower-scoring students (as rated by external listeners) would be more likely to show overconfidence simply because their comprehensibility is at the extreme (lower) end of the ability spectrum. However, as reported previously, students' comprehensibility scores, demonstrated a healthy range (1–9) both at T1 and T2 and were not limited to the lower end of the scale. Because various corrections for chance variation and bias in self-assessments have reduced but not entirely eliminated Dunning–Kruger effects observed in prior work (Burson et al., 2006; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Feld et al., 2017; Schlösser et al., 2013), the obtained negative associations, which are considered strong according to field-specific guidelines (Plonsky & Oswald, 2014), may thus have captured at least some aspects of the relationship compatible with the Dunning–Kruger effect. Students who were evaluated lower in comprehensibility by external listeners were those whose self-assessments demonstrated greater overconfidence.

5.2. Calibration of self-assessments over time

To investigate change in overconfidence, students were ranked by external listeners' ratings into performance quartiles (see Table 3): (a) high performers ($n = 26$), (b) upper-mid performers ($n = 27$), (c) lower-mid performers ($n = 27$), and (d) low performers ($n = 26$). A group ($4 \times$ time (2) ANOVA comparing overconfidence scores revealed a significant effect of group, $F(3, 102) = 73.89, p < .001, \eta_p^2 = 0.68$, and a significant two-way interaction, $F(3, 102) = 17.02, p < .001, \eta_p^2 = 0.34$, with no significant effect of time, $F(1, 102) = 0.05, p = .83, \eta_p^2 = 0.05$. According to Bonferroni-corrected comparisons (Fig. 3), all groups differed in their overconfidence scores at T1 ($p < .05$), but these scores significantly changed over time (towards self-assessment that was calibrated with listener ratings) for all groups ($p < .005$), except upper-mid performers ($p = .14$). Although high- and mid-level performers' overconfidence scores became similar

Table 3
Overconfidence by group ranked by external listeners' assessments.

Group	n	T1 (beginning of term)				T2 (end of term)		
		M	SD	95% CI	M	SD	95% CI	
1 High	26	-1.78	0.76	[-2.09, -1.47]	-0.76	0.89	[-1.12, -0.40]	
2 Upper-mid	27	-0.40	0.30	[-0.52, -0.28]	-0.10	1.01	[-0.50, 0.29]	
3 Lower-mid	27	0.50	0.37	[0.35, 0.65]	-0.07	0.98	[-0.46, 0.31]	
4 Low	26	1.68	0.51	[1.47, 1.89]	0.86	1.20	[0.37, 1.35]	

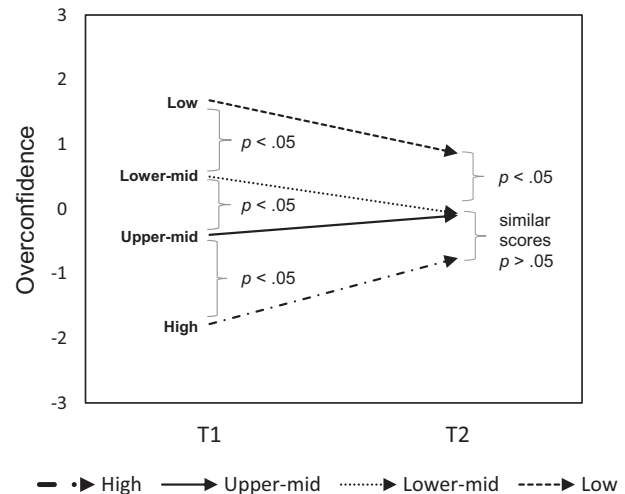


Fig. 3. Change in overconfidence (higher scores designate more overconfident ratings) by L2 comprehensibility performance (top, upper-mid, lower-mid, bottom performers) at T1 and T2.

at T2 ($p > .05$) and thus more aligned with external listeners' assessments, low performers remained significantly more overconfident than the other groups ($p < .01$).

Although a distance metric between self-assessments and external listeners' evaluations has been used to estimate speakers' overconfidence in prior research (e.g., Trofimovich et al., 2016; Hayes & Dunning, 1997), this measure may not successfully capture the complex nature of a calibration process over time (from T1 to T2). For example, if a speaker was overconfident at T1 and then became underconfident at T2, a relative difference in scores (T2 minus T1) could be quite large, which may misrepresent the magnitude of change relative to calibrated performance (the difference being zero). Therefore, to estimate absolute differences in the calibration process over time, we used a squared Euclidean distance metric (Table 4), which transformed differences between students' self-ratings and external listeners' assessments into absolute distances from the origin point (0), irrespective of positive (overconfidence) or negative (underestimation) values. These transformations were based on the raw ratings, not z scores.

To examine which group demonstrated changes in their Euclidean distance scores between T1 and T2, a set of Wilcoxon signed-rank tests (Bonferroni corrected $\alpha = 0.012$) were performed within each group (high, upper-mid, lower-mid, low performers). As illustrated graphically in Fig. 4, a statistically significant T1–T2 change occurred for high

Table 4
Squared euclidean distances by group ranked by external listeners' assessments.

Group	n	T1 (beginning of term)			T2 (end of term)		
		M	SD	95% CI	M	SD	95% CI
1 High	26	12.07	9.61	[8.18, 15.95]	5.81	1.14	[3.40, 8.10]
2 Upper-mid	27	1.25	1.04	[0.84, 1.66]	3.50	0.67	[1.98, 4.75]
3 Lower-mid	27	0.60	0.73	[0.31, 0.89]	3.46	0.66	[1.75, 4.49]
4 Low	26	6.22	4.44	[4.43, 8.02]	5.63	1.10	[1.53, 6.08]

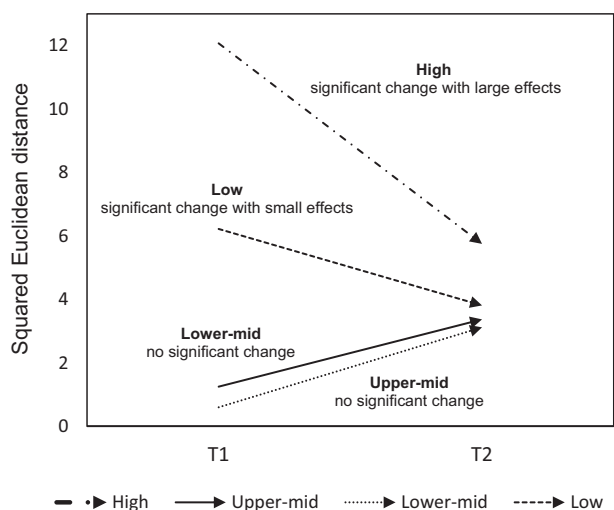


Fig. 4. Changes in squared Euclidean distance by group (high, upper-mid, lower-mid, low performers) at T1 and T2.

($p < .001$) and low ($p = .002$) performers, but not for mid-level performers ($p > .016$). These T1–T2 differences were associated with a large effect for high performers ($d = 0.88$), who became less likely to underestimate their comprehensibility, and with a medium effect for low performers ($d = 0.47$), who became less likely to overestimate their comprehensibility, relative to external listeners' ratings.

5.3. Calibration of self-assessments and students' individual difference profiles

Because the four comprehensibility-ranked groups (high, upper-mid, lower-mid, low) showed different calibration patterns between T1 and T2, the final analyses concerned the extent to which such group differences were associated with students' individual difference profiles. As discussed previously, there was variability in student experiences across the eight individual difference variables targeted in this study: (a) amount of L2 use in class, (b) number of hours per week spent on class preparation, (c) extra self-study at cram schools, (d) extra self-study through conversation activities, (e) Ideal L2 Self, (f) Ought-to L2 Self, (g) Anxiety, and (h) Enjoyment. Therefore, the goal of the final

Table 5
Factor analysis of experience, motivation, and emotion variables.

Measured variable	Factor 1: Promotional orientation	Factor 2: School-related practice	Factor 3: Extracurricular speaking practice
Speaking in class	.333	0.450	-.163
Preparation for class	0.218	-.729	0.135
Cram school	-.041	0.601	0.226
Conversation	0.284	0.337	0.609
Ideal L2 Self	.691	-.100	0.361
Ought-to L2 Self	-.0167	-.119	0.818
Anxiety	-0.797	-0.064	.086
Enjoyment	0.763	-0.135	-0.104

Note. All loadings > 0.6 highlighted in bold.

analysis was to examine whether the group differences observed among the high, upper/lower-mid, and low performers' calibration patterns were related to students' individual difference profiles, over and above any differences in their initial comprehensibility performance at T1.

Because seven students failed to fill in all items in the learner background questionnaire, a total of 99 students' data were included in the following analyses. To avoid multicollinearity (due to any overlap between student profile variables), a principal component analysis (PCA) with a direct oblimin rotation (for correlated variables) was conducted as a data reduction technique to uncover the latent categories underlying students' experience, motivation, and emotion characteristics. All variables entered into PCA involved ordinal data (6-point scales for motivation, anxiety, and enjoyment; 0–100% estimates for classroom language use), except for out-of-class language experience (extra study at cram schools and through conversational activities), which were binary (0, 1) data. Both ordinal and binary data are naturally ordered, and suitable for PCA (e.g., Gower, 1966, p. 332). Following Loewen and Gonulal's (2015) field-specific guidelines, the factorability of the dataset was considered relatively high, as shown by Bartlett's test of sphericity, $\chi^2 = 84.53, p < .001$, and the Kaiser-Meyer-Olkin measure of sampling adequacy (0.58). Using the maximum likelihood method, the model identified three factors with eigenvalues beyond 1.0, accounting for 56.37% of the total variance. A value of 0.60 was used as the threshold coefficient for practically significant factor loadings.

Each factor, summarized in Table 5, was interpreted as follows. Factor 1 was labeled "Promotional Orientation." Highlighting the interconnections between motivation and emotion (Teimouri, 2017), the three items with the highest loadings on this factor (in excess of 0.60) pertained to students' self-image as a desirable L2 user who experiences more positive emotions (more enjoyment) and fewer negative emotions (less anxiety). Factor 2 was labeled "School-Related Practice." The two highest factor loadings here captured the degree of students' preparedness for their English coursework and their participation in extra language practice at cram schools, where Japanese high-school students usually prepare for school exams with a long-term goal of passing university entrance exams. Factor 3 was labeled "Extracurricular Speaking Practice," because the two relevant items with factor loadings over 0.60 encompassed the extent to which students actively sought opportunities to practice L2 speaking skills with native and/or non-native speakers of English outside school instruction. Students who engaged in extracurricular speaking practice also appeared to have greater Ought-to L2 Self, ostensibly capturing individuals with a clear image of what their social networks or communities (e.g., friends, family members) expected them to achieve. This implies that these students may have felt external pressure to study English not only for exam-related purposes, but also for future professional and academic goals. Finally, the amount of students' speaking time inside English classes was not clearly associated with any of the above-mentioned factors. According to Kolmogorov-Smirnov tests, all factor scores were normally distributed ($p > .05$).

To examine the relationship between calibration in students' self-

Table 6
Correlations between self-assessment calibration and individual difference factors.

Factor score	Overconfidence scores (T2 minus T1)		Squared Euclidean distances (T2 minus T1)	
	<i>r</i>	<i>p</i>	<i>r</i>	<i>p</i>
Promotional focus	−0.127	.213	0.043	0.670
School-related practice	−0.093	0.363	0.061	0.549
Extracurricular practice	0.031	0.762	−0.256	0.010

assessment and their individual difference profiles, partial correlation analyses were performed using the entire sample of participants ($n = 99$). These analyses (Bonferroni corrected $\alpha = 0.025$) thus targeted the extent to which students' change in self-assessments, relative to external listeners' evaluations (T2–T1 overconfidence scores, T2–T1 squared Euclidean distance scores) could be tied to their individual differences in experience, emotion, and motivation. To control for pre-existing differences in initial levels of comprehensibility (both self- and other-assessed), students' T1 self-assessments of comprehensibility and T1 evaluations of students' comprehensibility by external listeners were partialled out from each correlation. As shown in Table 6, the amount of students' extracurricular speaking practice demonstrated a significant yet weak negative association with their T2–T1 squared Euclidean distance scores ($r = -0.26$, $p = .010$). This implied that students who better calibrated their self-assessments with the evaluations of external listeners were those who tended to engage in extracurricular practice activities on a more regular basis. More importantly, this relationship was independent of students' initial comprehensibility level.² For all other individual difference variables, however, the relationship between students' individual difference variables (from among those targeted in this study) and calibration of their comprehensibility self-assessments over time remained unclear.

6. Discussion

As a follow-up to the Trofimovich et al., 2016 study, which provided evidence for Dunning–Kruger effects in L2 speech assessment in a cross-sectional design, the current study aimed to extend these initial findings in a longitudinal investigation of L2 comprehensibility development for 106 Japanese EFL high school students over one academic semester (three months). Our analyses yielded four key findings. First, as in Trofimovich et al., 2016, EFL students demonstrated weak agreement between self- and other-perceptions of their L2 comprehensibility, especially at the onset of the term. Second, after engaging in one semester of EFL experience (seven 50-min EFL classes per week), students showed evidence of beginning to align their self-ratings with the assessments of external listeners. Third, this calibration depended on students' initial proficiency levels. High performers, who had initially

² As suggested by an anonymous reviewer, we also examined the opposite relationship, namely, whether students' individual difference profiles could explain some of the association between students' initial comprehensibility level and calibration of their self-assessments over time. A zero-order correlation between the students' initial comprehensibility scores and their T2–T1 squared Euclidean distance scores yielded no association ($r = -0.10$, $p = .30$); and a first-order correlation, with students' experience, emotion, and motivation scores partialled out, revealed no change in this relationship ($r = -0.08$, $p = .47$). This implied that any link between pre-existing differences in students' initial level of comprehensibility and calibration of their self-assessment over time was unrelated (at least in this dataset) to their individual difference profiles. However, we also call for future studies which can tease apart the impact of L2 speakers' initial proficiency and their individual difference profiles on the calibration process.

underestimated their comprehensibility, appeared to largely succeed at aligning their self-assessments with other-perceptions over time. In contrast, low performers, who had overestimated their comprehensibility initially, demonstrated a small amount of progress towards calibration. Finally, irrespective of students' initial comprehensibility level, those who had engaged in more extracurricular English practice tended to demonstrate greater change over time (T1 to T2) in the extent to which their self-assessments of comprehensibility aligned with those assigned by external listeners.

These findings are consistent with those reported in psychology (e.g., Dunning et al., 2004) and L2 assessment (e.g., Suzuki, 2015), showing that high and low performers frequently misjudge their ability. Echoing research findings in psychology (e.g., Hodges, Regehr, & Martin, 2001; Kruger & Dunning, 1999), the current results provided the first longitudinal evidence (over the span of three months of instructed language learning) for Dunning–Kruger effects in the context of L2 speech learning. Specifically, the findings showed that the extent of alignment in L2 speakers' self-assessment depends on speakers' initial L2 skill level. It is perhaps unsurprising that high performers can align their self-assessments with the evaluations by external listeners once they engage in more language learning experience over time, likely because this experience entails enhanced opportunities for speakers to become aware of their language ability relative to that of their peers. However, it seems that low performers may continue to demonstrate difficulty understanding how well they can perform in L2 speaking tasks over time.

A key question relevant to the present findings is which factors specifically can aid L2 speakers in noticing and minimizing the gap between their self-assessments and the evaluations by external observers (such as listeners, teachers, or examiners). In the context of this study, classroom instruction was unlikely to have been the catalyst for the change in self-assessments. One methodological strength of this study was that we had conducted classroom observations to document how English was taught in the high school EFL classrooms, finding no activity specifically devoted to self-assessment or to any type of awareness building regarding L2 speech or comprehensibility. Furthermore, none of the variables related to students' experience inside the classroom (i.e., Factor 2: School-Related Practice) showed any significant associations with the change in self-assessment.

Although weak in strength, the obtained correlation coefficients implied a mediating role for extracurricular speaking practice in students' calibration process: Those who spent time learning English at cram schools and through voluntary conversation activities beyond the school curriculum may have engaged in more accurate self-assessments of their comprehensibility, relative to the judgments of external listeners. The findings here echo previous literature which has shown that students' extra efforts to seek opportunities to use the target language outside classrooms could be instrumental to their success, particularly in foreign language settings (Muñoz, 2014; Muñoz, Cadierno, & Casas, 2018; Saito & Hanzawa, 2016). It is these conversational opportunities, where students can receive and benefit from interlocutors' corrective feedback, that are likely of great acquisitional value, especially when communication breakdowns take place due to language-related problems leading to negotiation for meaning (Mackey, 2012; for longitudinal evidence, see Saito & Akiyama, 2017).

It is noteworthy that the extracurricular practice factor was strongly tied to students' motivation to achieve what their external networks (e.g., friends, family members) likely expected of them (Ought-to L2 Self) rather than what students visualized for themselves (Ideal L2 Self). Although Ideal L2 Self may be the motivational factor responsible for encouraging L2 speakers to practice the target language, often through emotionally positive experiences, leading to measurable learning gains (e.g., Teimouri, 2017), our findings suggest that it is rather Ought-to L2 Self that might affect the calibration process between self- and other-assessments in foreign language contexts like Japan. This is arguably because, in foreign language settings, learning is more strongly driven

by the expectations imposed on learners rather than by their own future images (Li, 2014). We agree with Nagle (2018) that, in order to clearly understand the link between motivational orientations, self-assessments, and language learning, researchers must engage in both quantitative and qualitative work.

With respect to students' experience inside classrooms, the current results did not reveal its contribution to change in students' self-assessments. In classroom observations, there was no evidence that teachers implemented any specific awareness-raising activities; and the three months of EFL instruction captured in this study was not the first or only classroom-based experience for students. Presumably, prior instruction could have provided students with ample opportunities to refine their self-assessments, yet these experiences did not lead to accurate self-ratings. What likely mattered was the explicit nature of self-assessments built into the research, where students rated their comprehensibility twice during the term (test–retest effect), which corresponds to a pedagogical intervention not systematically practiced in many classrooms (Butler & Lee, 2010). If faulty self-assessments can be remedied with minimal intervention (through repeated assessments), then this is good news for L2 speakers and their teachers.

However, such path to accurate self-assessment could be observed in this dataset only among high performers. These students are assumed to have sufficiently adequate, advanced L2 skills in order to analyze and compare themselves to other, often less proficient students in an effective, meaningful fashion. Comparably successful calibration may not occur for low performers who likely lack enough experience and competence to engage in meta-comparison behaviors. In this study, low performing students did spend additional time in extracurricular practice throughout the project, an activity which we found to be conducive to the development of self-assessment accuracy. However, the extent of these students' assessment calibration over time was rather small.

The next relevant question, one with much pedagogical relevance, is how less experienced and less proficient students could be helped, in particular, with correctly evaluating their L2 speaking ability, including L2 comprehensibility, given that they are otherwise unlikely to become aware of the gap between self- and other-assessed performance through increasing their exposure to the target language. There has been some evidence that inexperienced L2 speakers can benefit from awareness-raising activities which explicitly direct speakers' attention to differences between their own assessments and external evaluations (Butler & Lee, 2010; de Saint Léger, 2009). In addition, prior research has also shown that L2 speakers can improve their self-assessment accuracy when they participate in relatively easy, simple self- and peer-assessment activities (Hayes & Dunning, 1997) while receiving immediate feedback from teachers (Butler & Lee, 2010). The effectiveness of these and other instructional treatments for the development of L2 speakers' self-assessment of their speaking skills must be revisited in future research.

7. Limitations and future research

Although this study explored Dunning–Kruger effects in L2 speech learning using both cross-sectional and longitudinal analyses, the findings must nevertheless be interpreted as tentative. One reason is that we did not use refined measures tapping into various types of linguistic experience for participants and thus failed to capture the complexity of their individual linguistic repertoires. Indeed, participants' experiential profiles were relatively crude, with all measures based on self-reports. In addition, we could not ascertain possible links between individual participants' specific experiences within language classrooms and their self-assessment performance. It is possible that different L2 speakers respond differently to essentially the same classroom instruction, revealing aptitude–treatment interaction effects (Vatz, Tare, Jackson, & Dougherty, 2013) which we failed to capture.

We only focused on one global dimension of L2 speaking (comprehensibility), as assessed through Likert-type ratings in a single task.

There was also a mismatch between what participants were asked to evaluate (i.e., their own comprehensibility in general) and what external listeners were asked to evaluate (i.e., speakers' comprehensibility elicited from a particular task), implying that L2 speakers' and external listeners' assessments may have reflected task-specific performance to different degrees. As argued by De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2012), various dimensions of L2 speaking proficiency—and by extension accuracy of L2 speakers' self-assessment of these dimensions—depend crucially on the specific measures and tasks employed to target these dimensions. Future research into L2 speakers' self-assessments should use other, more valid measures to reduce the possibility that at least some over- or under-confidence in L2 speakers' judgment occurs due to the regression-to-the-mean effect (Krueger & Mueller, 2002), particularly for speakers at extreme ends of the ability continuum. One useful measure might include direct magnitude estimation, where speakers (and external listeners) compare the target speech sample with a given reference item (Brennan, Ryan, & Dawson, 1975; Munro, 2018). To obtain deeper insights into Dunning-Kruger effects, future studies may also need to adopt quantitative and qualitative analyses in a complementary fashion. For example, it would be interesting to supplement quantitative modelling with retrospective interviews with high- and low-ability L2 speakers to understand meta-cognitive processes that they engage in during repeated self- and peer-assessments over time.

In future L2-focused work, researchers should also separate meta-cognitive contributions to self-assessments from statistical or methodological biases, either through separate measures of metacognition and response bias (Burson et al., 2006; Kruger & Dunning, 1999) or through correction for response bias (Feld et al., 2017; Krueger & Mueller, 2002; Kruger & Dunning, 2002). Last but not least, researchers should clarify potential sources of inaccurate L2 self-assessment or lack of calibration over time. As a construct, comprehensibility certainly qualifies as an ill-defined, fuzzy, and subjective domain, where self-assessment is at risk of being inaccurate (Burson et al., 2006; Hayes & Dunning, 1997). Indeed, speech is a complex phenomenon involving multiple factors (e.g., prosody, lexis, syntax, affect, emotion). Similarly, L2 speakers can be understood by their interlocutors despite having a noticeable accent and through interlocutors having access to shared context or through speakers making use of various strategies (e.g., gesturing, avoidance, circumlocution). It would thus be important to understand whether and to what degree inaccurate self-assessments in a complex domain like comprehensibility, where “success” is based on multiple, intertwined performance dimensions, are driven by lack of metacognitive knowledge (e.g., Dunning et al., 2004), reflect L2 speakers' attribution of success to others (e.g., Fussell & Krauss, 1992), arise as a consequence of specific entrenched, preconceived self-views (e.g., Critcher & Dunning, 2009), or encapsulate an erroneous decision strategy (e.g., Dunning, 2019).

8. Conclusion

Despite several shortcomings, the current findings yielded valuable insights into classroom language learners' self-assessment behaviors. First, high and low performing L2 learners differed in extent to which they calibrate their self-assessments of comprehensibility with external listeners' assessments, in the absence of any focused awareness-raising activities. And second, the extent to which learners engaged in additional practice of the target language, especially outside regular instruction, was positively associated with learners' self-assessment accuracy over time. These findings motivate future work examining the effectiveness of various instructional treatments designed to help L2 learners to calibrate their self-assessments with external evaluations in various assessment contexts.

Declaration of competing interest

We have no conflict of interest in this current manuscript.

Acknowledgments

The project was funded by a Grant-in-Aid for Scientific Research in Japan 16H03455 (awarded to the last author) and an Arnold Bentley New Initiatives Fund (awarded to the first author).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.lindif.2020.101849>.

References

- ALC Press Inc. (2017). Telephone Standard Speaking Test. Retrieved from https://tsst.alc.co.jp/tsst/e_index.html.
- Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing*, *33*, 1–27.
- Brennan, E. M., Ryan, E. B., & Dawson, W. E. (1975). Scaling of apparent accentness by magnitude estimation and sensory modality matching. *Journal of Psycholinguistic Research*, *4*, 27–36.
- Burson, K. A., Larrick, R. P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: How perceptions of difficulty drive miscalibration in relative comparisons. *Journal of Personality and Social Psychology*, *90*, 60–77.
- Butler, Y., & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing*, *27*, 5–31.
- Collins, L., Trofimovich, P., White, J., Cardoso, W., & Horst, M. (2009). Some input on the easy/difficult grammar question: An empirical study. *The Modern Language Journal*, *93*(3), 336–353. <https://doi.org/10.1111/j.1540-4781.2009.00894.x>.
- Critcher, C. R., & Dunning, D. (2009). How chronic self-views influence (and mislead) self-assessments of task performance: Self-views shape bottom-up experiences with the task. *Journal of Personality and Social Psychology*, *97*, 931–945.
- Crowther, D., Trofimovich, P., Saito, K., & Isaacs, T. (2015). Second language comprehensibility revisited: Investigating the effects of learner background. *TESOL Quarterly*, *49*, 814–837.
- De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition*, *34*, 5–34.
- de Saint Léger, D. (2009). Self-assessment of speaking skills and participation in a foreign language class. *Foreign Language Annals*, *42*, 158–178.
- Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam: John Benjamins.
- Dewaele, J.-M., & MacIntyre, P. D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching*, *4*, 237–274.
- Dewaele, J. M., Witney, J., Saito, K., & Dewaele, L. (2018). Foreign language enjoyment and anxiety: The effect of teacher and learner variables. *Language Teaching Research*, *22*, 676–697.
- Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. London: Routledge.
- Dragojevic, M., & Giles, H. (2016). I don't like you because you're hard to understand: The role of processing fluency in the language attitude process. *Human Communication Research*, *42*(3), 396–420.
- Dunning, D. (2011). The Dunning-Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson, & M. P. Zanna (Vol. Eds.), *Advances in experimental social psychology*. Vol. 44. *Advances in experimental social psychology* (pp. 247–296). New York, NY: Academic Press.
- Dunning, D. (2019). The best option illusion in self and social assessment. *Self and Identity*, *18*, 349–362.
- 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.
- Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*, 98–121.
- EIKEN Foundation of Japan. Retrieved from <https://www.eiken.or.jp/eiken/en/research/comparison-table.html> (Last viewed September, 2017).
- Ellis, R. (1997). *Second language acquisition*. Oxford, UK: Oxford University Press.
- Feld, J., Sauermaun, J., & de Grip, A. (2017). Estimating the relationship between skill and overconfidence. *Journal of Behavioral and Experimental Economics*, *68*, 18–24.
- Fussell, S. R., & Krauss, R. M. (1992). Coordination of knowledge in communication: Effects of speakers' assumptions about what others know. *Journal of Personality and Social Psychology*, *62*, 378–391.
- Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, *53*, 325–338.
- Hayes, A. F., & Dunning, D. (1997). Construal processes and trait ambiguity: Implications for self-peer agreement in personality judgment. *Journal of Personality and Social Psychology*, *72*, 664.
- Hodges, B., Regehr, G., & Martin, D. (2001). Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine*, *76*, S87–S89.
- Horwitz, E. K. (2017). On the misreading of Horwitz, Horwitz, and Cope (1986) and the need to balance anxiety research and the experiences of anxious language learners. In C. Gkonou, M. Daubney, & J.-M. Dewaele (Eds.), *New insights into language anxiety: Theory, research and educational implications* (pp. 31–47). Bristol: Multilingual Matters.
- Kang, O., Thomson, R. I., & Moran, M. (2018). Empirical approaches to measuring the intelligibility of different varieties of English in predicting listener comprehension. *Language Learning*, *68*, 115–146.
- Kennedy, S. (2009). L2 proficiency: Measuring the intelligibility of words and extended speech. In A. Benati (Ed.), *Issues in second language proficiency* (pp. 132–144). London: Continuum.
- 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.
- Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*, 180–188.
- Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*, 1121–1134.
- Kruger, J., & Dunning, D. (2002). Unskilled and unaware—But why? A reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology*, *82*, 189–192.
- 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.
- Lev-Ari, S., & Keyser, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of Experimental Social Psychology*, *46*(6), 1093–1096.
- Li, Q. (2014). Differences in the motivation of Chinese learners of English in a foreign and second language context. *System*, *42*, 451–461.
- Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research*. New York: Routledge.
- Ludwig, A., & Mora, J. C. (2017). Processing time and comprehensibility judgments in non-native listeners' perception of L2 speech. *Journal of Second Language Pronunciation*, *3*(2), 167–198.
- Mackey, A. (2012). *Input, interaction, and corrective feedback in L2 learning*. Oxford, UK: Oxford University Press.
- Muñoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics*, *35*, 463–482.
- Muñoz, C., Cadierno, T., & Casas, I. (2018). Different starting points for English language learning: A comparative study of Danish and Spanish young learners. *Language Learning*, *68*, 1076–1109. <https://doi.org/10.1111/lang.12309>.
- Munro, M. J. (2018). Dimensions of pronunciation. In O. Kang, R. I. Thomson, & J. M. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 413–431). New York: Routledge.
- Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*, 289–306.
- Nagle, C. (2018). Motivation, comprehensibility, and accentness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal*, *102*, 199–217.
- Nagle, C., Trofimovich, P., & Bergeron, A. (2019). Toward a dynamic view of second language comprehensibility. *Studies in Second Language Acquisition*, *41*, 647–672.
- Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.
- Saito, K., & Akiyama, Y. (2017). Video-based interaction, negotiation for comprehensibility, and second language speech learning: A longitudinal study. *Language Learning*, *67*, 43–74.
- Saito, K., & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied Psycholinguistics*, *37*, 813–840.
- Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*, 652–708.
- Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentness: A validation and generalization study. *Applied Linguistics*, *38*, 439–462.
- Saito, K., Dewaele, J. M., Abe, M., & In'nami, Y. (2018). Motivation, emotion, learning experience, and second language comprehensibility development in classroom settings: A cross-sectional and longitudinal study. *Language Learning*, *68*(3), 709–743. <https://doi.org/10.1111/lang.12297>.
- Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the “signal extraction” counterexplanation for the Dunning-Kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, *39*, 85–100.
- Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 1–32). Cambridge, UK: Cambridge University Press.
- Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing*, *32*, 63–81.
- Taguchi, T., Magid, M., & Papi, M. (2009). The L2 motivational self system among Japanese, Chinese and Iranian learners of English: A comparative study. In Z. Dörnyei, & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 66–97). Clevedon: Multilingual Matters.
- Teimouri, Y. (2017). L2 selves, emotions, and motivated behaviors. *Studies in Second*

- Language Acquisition*, 394, 681–709.
- 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(1), 122–140. <https://doi.org/10.1017/S1366728914000832>.
- Vatz, K., Tare, M., Jackson, S. R., & Doughty, C. J. (2013). Aptitude-treatment interaction studies in second language acquisition: Findings and methodology. In G. Granena, & M. Long (Eds.). *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 273–292). Amsterdam, The Netherlands: John Benjamins.