System 123 (2024) 103352

Contents lists available at ScienceDirect

System

journal homepage: www.elsevier.com/locate/system

What influences the comprehensibility of L2 writers' opinion texts by L2 readers? Interactions between textual characteristics and readers' profiles

Miyuki Sasaki [a,*], Yui Suzukida [b], Kotaro Takizawa [a], Kazuya Saito [c]

[a] *Waseda University, Japan*
[b] *Juntendo University, Japan*
[c] *University College London, UK*

A B S T R A C T

This study explored how L2 readers perceive the comprehensibility (i.e., ease of understanding) of L2 opinion texts in terms of reader profiles and text characteristics. We investigated how 100 Japanese learners of L2 English evaluated the comprehensibility of 16 variants of two opinion essays, which differed in lexico-grammatical accuracy, coherence, and rhetorical organization (inductive vs. deductive structure). To delve into the evaluators' backgrounds, exploratory factor analysis was conducted with a view to streamlining background variables. These variables were then compared through a *t*-test of two groups of readers identified by cluster analysis based on their degree of leniency or strictness. We further identified the influence of specific linguistic features on their evaluations by conducting linear mixed-effects analyses. Our findings reveal that: (1) Some L2 readers were significantly stricter than the L1 readers, whereas others were more lenient; (2) The lenient readers typically began learning English earlier, used it more extensively outside academic settings, and engaged more frequently in online communication compared to their stricter counterparts; and (3) Though disrupted sequences of ideas (coherence errors) were universally detrimental to comprehensibility, lexico-grammatical errors significantly impacted only the strict readers, not the lenient ones.

1. Introduction

As many researchers have noted, English has acquired the status not only of lingua franca but also of "international gatekeeper" (Pennycook, 2017, p. 13), and academic texts written in English as an additional language are no exception (e.g., Flowerdew, 2022). Given that over 70% of English users are not L1 speakers (Statista, 2023), we need to examine how the English academic texts of second language (L2) writers are perceived by multilingual readers.

In L2 writing research, learners' academic writing skills have tended to be evaluated by L1 English speakers or by trained L1 or L2 English-speaking teachers (e.g., Zhang & Cheng, 2021). This is most likely due to the long-held belief that the ultimate goal of learning English is to speak and write like an L1 English speaker, which many researchers such as Ortega (2018, p. 66) now dismiss as "nativespeakerism." While such L1 speaker normativism may be essential in advanced academic contexts, including in studies published in international journals (e.g., Flowerdew, 2022), college-level academic L2 writing is generally taught in classrooms, and its ultimate goals are not as demanding (e.g., Ministry of Education, Culture, Sports, Science and Technology, Japan, 2023). Furthermore, although college-level writing forms a sub-category of academic writing, it may have a wider audience and thus needs to accommodate to various multilingual contexts (Ferris & Hedgcock, 2023). Among various types of L2 English texts, in this study, we target texts written by college-level students of English as a second or foreign language (ESL or EFL). To our knowledge, few studies of L2 writing,

* Corresponding author. 1-6-1-16, Nishiwaseda, Shinjukuku, Tokyo, 169-8050, Japan.
E-mail address: miyuki.sasaki@waseda.jp (M. Sasaki).

https://doi.org/10.1016/j.system.2024.103352
Received 10 March 2023; Received in revised form 20 May 2024; Accepted 24 May 2024
Available online 27 May 2024

whether academic or non-academic, have looked at how non-L1 English speakers evaluate EFL/ESL learners' writing in terms of comprehensibility (operationally defined as "ease of understanding; " see Saito et al., 2019, p. 1133).

On the other hand, when the mode changes from writing to speaking, the research landscape is markedly different, particularly in terms of evaluation criteria relevant to the construct of *comprehensibility* as well as the evaluators' background. First, in L2 speech evaluation, due to its strong association with pronunciation, along with intelligibility, a comparable yet distinct phonological aspect of speech, comprehensibility has long been regarded as one of the pivotal determinant of L2 speech quality, leading to extensive investigations (e.g., Crowther et al., 2022). More specifically, since Munro and Derwing's (1995) seminal study, the comprehensibility of L2 speech has been explored in relation to other features that might influence listeners' perceptions. These features encompass sound-specific elements such as "segmental and word stress errors, intonation, and speech rate" (Foote & Trofimovich, 2018, p. 253) as well as broader language-related aspects such as lexical sophistication (e.g., Saito & Shintani, 2016), fluency, and grammatical accuracy (e.g., Saito & Shintani, 2016). Another distinction to be made between writing and speaking in evaluating L2 text is listeners' background. In speech studies, even though listeners are typically L1 speakers of the target language, they are often asked to make impressionistic judgments of how effortlessly they can understand L2 learners' speech (Trofimovich & Isaacs, 2012; Nagle, 2018). This contrasts with L2 writing assessments, where there is greater emphasis on evaluators' qualification and training (Ferris & Hedgcock, 2023). In some L2 comprehensibility studies, listeners' qualifications (such as teaching experience) are generally required for purposes such as developing a valid rating scale tailored to specific testing and teaching contexts (e.g., Isaacs et al., 2018). More commonly, however, individual variables such as familiarity with the speakers' L1 are employed to investigate variance in L2 speech comprehensibility (Crowther et al., 2022).

Variables such as familiarity with speakers' L1 eventually led some researchers to study L2 listeners whose L1 matched that of the speakers of the texts they listened to (e.g., Foote & Trofimovich, 2018; Ludwig & Mora, 2017; O'Brien, 2014, 2016). Although these researchers predominantly used L1-speaking listeners as a benchmark in comparing their evaluations with the reactions of participating L2 listeners of L2 speech, this may amount to an implicit endorsement of native-speaker norms. Nevertheless, these studies were among the first to explore the involvement of multilingual listeners in L2 speech research. Subsequently, Saito and Shintani's study (2016) was epoch-making because it was conducted from a non-deficit view of L2 users (i.e., learners' existing skills can be resources), which better accommodates the multilingual perspective now widely shared in applied linguistics (e.g., Douglas Fir Group, 2016). In their study, Saito and Shintani took Canadian English L1 speakers not as the model they should refer to but as one monolingual group as opposed to a multilingual Singaporean group for the purpose of comparison. They found that the Singaporeans gave more lenient scores than the Canadians for comprehensibility of English speech delivered by Japanese speakers with various levels of English proficiency.

These studies represent a significant advance in that they involved multilingual listeners evaluating L2 texts. However, if the outcomes of evaluations are determined by interactions between assessment behaviors and the diverse backgrounds of the evaluators, compared with studies targeting L1-speaking listeners, these studies are still limited in terms of a variety of listeners characteristics (e.g., Foote & Trofimovich, 2018). More specifically, in early L1-matched studies, the examination of the backgrounds of the L2 listeners was confined to their L1. Even in Saito and Shintani (2016), the variable was whether the listeners were monolingual or multilingual, which still poses a limitation. Consequently, the exploration of the relationship between the comprehensibility of L2 texts by multilingual users and the various profiles of evaluators was a first and novel contribution by Saito et al. (2019). Details of this study are used as the model framework for the present study and are elaborated upon in Section 2.2.

Given such knowledge accumulation in L2 speech studies, the paucity of L2 writing studies investigating L2 users evaluating comprehensibility of L2 writers' texts is regrettable, especially in an increasingly multilingual world. In response, building on the framework established by Saito et al. (2019), we conducted a pioneering study exploring L2 readers' evaluations of L2 writing and how their diverse profiles (e.g., L2 use outside the classroom) affected their rating.

2. Background

2.1. Evaluation of comprehensibility in L2 academic writing

We saw in the previous section that in order to understand the multifaceted nature of L2 speech evaluation, it is crucial to consider linguistic features of the L2 speech that may affect listeners' perceptions as well as the listeners' own background. Following the same logic, to investigate the multi-faceted nature of L2 writing comprehensibility, we will need to ask two questions: (A) What linguistic features are most likely to affect readers' perceptions of L2 text comprehensibility? and (B) How is readers' own background related to their comprehension of L2 texts? Before we ask these two questions, however, we must first ask a more fundamental question, namely what "comprehensibility" in L2 writing consists of.

2.1.1. Comprehensibility in L2 writing

Comprehensibility is a multifaceted construct in second language acquisition (SLA) research. As described above, while investigation of this construct has been extensive in the L2 speech research, its application as a central trait in evaluating L2 academic writing has not been similarly foregrounded (e.g., Ferris & Hedgcock, 2023). Instead, comprehensibility in writing has often been subsumed as one of the analytical subconstructs under the same label (i.e., "comprehensibility") (e.g., Kuiken & Vedder, 2017 or under similar labels ("clarity of meaning" in Kobayashi & Rinnert, 1996; "readability" in Zhang et al., 2022). As described in Section 2.3, because we selected Kobayashi and Rinnert (1996) as our benchmark study, we decided to use "clarity of meaning," one of the analytical criteria used by these researchers as a construct equivalent to "comprehensibility," as the central construct of our investigation due to its

similarity to our operational definition of "comprehensibility," i.e., "ease of understanding." To our knowledge, no study to date has explored Questions A and B above in order to better understand the nature of comprehensibility as the focal construct. Even in Kobayashi and Rinnert, the construct of "clarity of meaning" is one of the subconstructs used to explain the targeted construct of "cultural rhetorical patterns."

2.1.2. Constructs that may affect text comprehensibility for L2 readers

To identify possible characteristics of L2 texts that may affect L2 readers' comprehension, we need to start with the types of assessment criteria widely used in L2 academic writing at college level, our target in this study. In general, assessment of the quality of academic L2 writing at this level involves various constructs, as represented by authoritative tests of English as a second or foreign language (ESL/EFL) such as TOEFL iBT and IELTS. Of these, some of the most frequently mentioned constructs include organization, lexico-grammatical knowledge, and coherence, which have been extracted from a substantial number of needs analyses supported by rigorous validation processes (e.g., Chapelle et al., 2008). In fact, as Kobayashi and Rinnert (1996) argued, the three constructs of organization, lexico-grammatical knowledge, and coherence "correspond to the three frequently evaluated traits" (p. 402; see also Ferris & Hedgcock, 2023).

That said, we are aware that in both ESL and EFL classrooms, evaluators tend to apply the L1 English norms of one or other Inner Circle variety (Ferris & Hedgcock, 2023; Hyland, 2019). It may therefore be inappropriate to impose such norms on L2 readers who are not L1-English speakers. In fact, given the recent rise in conceptualizations of English as a Lingua Franca (ELF) and World Englishes, a growing body of research has been moving away from such a normative approach. Among them is the approach based on the WrELFA academic ELF corpus compiled by Mauranen and colleagues (2020). Since in rating "good texts written in non-standard English" (p. 71) "success is determined by the achievement of communication" (Mauranen, 2018, p. 113), traits such as accuracy may not be crucial.

A similar sentiment is echoed by ELF researchers such as Carey (2013) and Mur-Dueñas (2013) as well as corpus linguistics researchers such as Rozycki and Johnson (2013), who investigated multilingual writers' engineering research papers. However, despite these researchers advocating the assessment of multilingual writers' texts on the basis of comprehensibility and the spread of the notion of World Englishes, a majority of gatekeepers within academia remain "very conservative, relying heavily on models of native-speaker English as the linguistic standard" (Flowerdew, 2022, p. 580). This view is supported by empirical studies that examined how much effort English as additional language writers must invest to make their manuscripts fit a prescriptive "standardized model" (Flowerdew, 2022, p. 572) if they are to publish in international journals (e.g., Flowerdew & Wang, 2016). Given this background, although the written L2 English proficiency of our targeted participants fell far short of publishable quality, we decided to retain the three frequently evaluated traits of organization, lexico-grammatical knowledge, and coherence in L2 writing, the academic contexts we wished to investigate.

2.1.3. Relationship between L2 readers' L2 text comprehension and their background

As mentioned above, we chose Kobayashi and Rinnert (1996) as our benchmark study because it provided us with appropriate background in our search for the central construct of *comprehensibility* as well as for the three variables that may affect our participants' comprehensibility ratings. However, to determine what specific background characteristics we should investigate, we used Saito et al. (2019) as a model study because it was one of the first studies to explore the interplay between L2 listeners' comprehension of L2 speech and their various individual characteristics (Section 1; see also Section 2.2). In that study, participants' characteristics were selected as variables in a data-driven manner, and overlapping categories were collapsed through factor analysis.

Having limited our focus to studies of academic L2 texts, we had no prior studies that might guide our examination of the relationship between L2 readers' evaluation behaviors and their backgrounds, except for Kobayashi and Rinnert (1996). However, compared with Saito et al. (2019), Kobayashi and Rinnert only used the participants' L1s, academic status (teacher vs. student), and amount of L2 writing instruction as their participants' characteristics, whereas we wished to use a greater variety of characteristics for our exploration. Moreover, even though some researchers have investigated how different types of multilingual readers react to various aspects of academic L2 texts, most such readers were L1-speaking teachers or testers who tended to apply conventional normative "models of English" (Flowerdew, 2022, p. 571), which makes most of these studies largely irrelevant to ours.

Most relevant to our study are studies in which at least some participating readers were multilingual users, and the relationship between their rating behaviors and their backgrounds was investigated, even though the researchers' main focus was to explore the rating behaviors of L1-speaking teachers or testers. For example, Connor-Linton (1995) had half a group of L1 English and half a group of L1 Japanese instructors rate holistically and the other half of each group rate analytically the same ten English compositions written by adult Japanese learners and provide reasons for their ratings. He found that both groups of instructors gave similar overall ratings but that the American and Japanese instructors paid attention to different aspects of the writing (e.g., grammar, coherence). More recently, Hyland and Anan (2006) also compared L1 English and L1 Japanese teachers of English ($n = 16$ for each group) and found that the L1 Japanese teachers were stricter and paid greater attention to grammatical features than did the L1 English teachers, who cared more about overall comprehensibility. Finally, Shi (2001) conducted a similar study comparing L1 English and L1 Chinese teachers of English and found that while the two groups' holistic ratings were similar, their paths to reaching their final decisions were significantly different.

Several studies have explored the relationship between the background of multilingual readers who were neither teachers nor testers along with their reactions to L2 texts. For example, Santos (1988) asked 178 professors at US universities to rate two English L2 compositions in terms of content (holistic impression, development, sophistication) and language use (comprehensibility, acceptability, and readers' degree of resulting irritation, if any). Of the 178 professors, 96 were humanities majors and 82 were physical

sciences majors, and 144 were L1-English speakers while 34 were non-L1 speakers, with ages ranging from 27 to 77. Focusing on the 34 non-L1-English-speaking professors, Santos found that they gave significantly lower acceptability scores than did their L1-English-speaking counterparts. In another study, Kobayashi (1992) asked participants with L1 English (n = 145) and L1 Japanese/L2 English (n = 142) working in ESL contexts in the US and the UK to rate an L2 English composition written by an L1 Japanese writer for grammaticality, clarity of meaning, naturalness, and organization. However, instead of using the L1 speakers as the norm, the researcher aimed to capture interactions between the participants' L1s. Results revealed that the L1 Japanese/L2 English readers were generally more lenient over grammaticality than the L1 English readers. In contrast, the L1 English professors and graduate students tended to be more lenient than their L1 Japanese counterparts over clarity of meaning and organization whereas the reverse was true at the undergraduate level.

As mentioned above, Kobayashi and Rinnert is one of very few studies that have investigated L2 users' rating behaviors of L2 texts while also exploring the effects of their various backgrounds. Participants in that study were divided into four groups: i) Japanese L1 students with no specific L2 English writing instruction in their respective university (inexperienced students); ii) Japanese L1 students with at least one semester of L2 English writing instruction (experienced students); iii) Japanese L1 teachers of English (Japanese teachers); and iv) L1 English-speaking teachers (English teachers). Of these, the first three groups were multilingual readers. The researchers had all four groups of readers evaluate 16 L2 English compositions manipulated in terms of topic, rhetorical organization, lexico-grammatical features, and coherence. The results that are particularly relevant to our own study were that: (1) the inexperienced students gave significantly higher overall scores to the inductive (Japanese) L2 texts than did the English teachers; (2) the Japanese teachers gave significantly higher scores to the Introduction in the inductive L2 texts; and (3) the English teachers gave significantly higher scores to the Conclusion in the deductive L2 texts compared to the inexperienced Japanese students. In short, both L1 background and length of L2 writing instruction were found to influence the readers' evaluation behaviors.

2.2. *Saito et al. (2019) as the model framework*

Saito et al. (2019) investigated: (1) whether there were any differences in rating patterns between various L2 listeners; (2) what speech features these listeners attended to when evaluating speech comprehensibility; and (3) what profiles (e.g., L1, age, duration and quality of English learning and use, and metacognition about L2 speech) affected the comprehensibility ratings of these listeners. The researchers recruited 110 multilingual adult users of English and had them rate the comprehensibility of the English speech of 50 Japanese L1 speakers. The speech samples were linguistically analyzed in terms of type and number of errors (e.g., accentedness, phonological accuracy) as well as fluency, and the L2 listeners were divided into L2 lenient and L2 strict groups according to their comprehensibility judgements. Results indicated that: (1) L1 listeners were more strongly affected by pronunciation accuracy, whereas L2 listeners' evaluations were more equally affected by different linguistic features such as pronunciation, fluency, and lexico-grammatical error features; (2) the L2 lenient group tended to pay more attention to "lexical appropriateness and fluency in their judgment" (p. 1146) compared to the L2 strict group; and (3) compared to the L2 strict listeners, the L2 lenient listeners tended to have "higher levels of awareness of the importance of comprehensibility for communication (metacognition), regularly used L2 English in professional settings (experience), and had L1s more linguistically close to the target speech samples" (p. 1134).

2.3. *Kobayashi and Rinnert (1996) as the benchmark L2 writing study*

As mentioned above, Kobayashi and Rinnert (1996) provided our study with two important tools. The first was a measure of the construct of *comprehensibility* for L2 texts. Second, we are grateful to Kobayashi and Rinnert for allowing us to adapt their 16 written English academic texts as our stimulus materials because they were well suited to our purpose (see Section 3.2).

In addition, we kept the variable of writers' L1 (i.e., Japanese) constant as L1-L2 distance was a significant factor in Saito et al. (2019). Since ours was to be the first study of this kind, we opted to limit ourselves to just one potentially powerful explanatory variable, namely the participants' L1, unlike Saito et al., whose 110 L2 speakers spoke eight typologically different languages.

Based on these considerations, we formulated the following research questions.

1. What linguistic features (e.g., rhetorical patterns, lexico-grammatical errors, disrupted sequences of ideas) of opinion texts most strongly influence L2 readers' evaluations of comprehensibility?
2. How was readers' background related to their comprehension of L2 texts?

3. Method

Following the framework in Saito et al. (2019), we first administered a background questionnaire to explore the readers' profiles. Participants then engaged in reading and evaluating all 16 L2 English texts using a 10-point comprehensibility scale. This step was crucial for associating ratings with the three above-mentioned inherent linguistic features of the texts, thus providing an empirical basis for comprehensibility evaluation. Factor analysis was employed to reduce reader variables and thus mitigate multicollinearity, which enabled us to identify distinctive features crucial to differentiating between lenient and strict readers effectively. This approach led us to draw nuanced conclusions regarding the impact of varied reader profiles on comprehensibility assessments. Subsequent sections offer detailed descriptions of each procedural step.

3.1. Participants

Data collection was conducted in November 2021. Due to the COVID-19 pandemic, all procedures were carried out online (through Zoom). Instructions were given in the readers' L1 (Japanese). One hundred university students participated in the study. They were recruited online from classes related to linguistics in the home universities of the first, second, and third authors. They were asked to participate in a 50- to 70-minute in-class experiment (those uninterested in taking part could leave the room without consequences). Of the 100 participants, 14 were graduates majoring in social sciences, and 86 were undergraduates majoring in social sciences and science. Of these, 28 had taken TOEFL ITP during the previous year, with an average score of 540.5, indicating that their English was at the CEFR B1 (lower -independent) level. In these particular universities, all departments require English skills for admission, and many departments also require (or make optional) English-medium classes.

To set the baseline, we also recruited 15 L1 English speakers (6 Male, 8 Female; Mean age = 20.6, SD = 1.68, Range 19–24) based in England through Prolific (www.prolific.com), an online platform designed to facilitate the recruitment and management of participants for online research. The second author gave the participants appropriate instructions and training in the evaluation rubric. However, they were not informed that the texts they were to evaluate had been intentionally modified or written by L2 users of English.

3.2. Evaluation criterion and stimulus preparation

To answer RQ1 and RQ2, we prepared materials that would allow us to identify text characteristics that may potentially influence the participants' comprehensibility judgements. First, as mentioned above, we used *clarity of meaning* (Kobayashi & Rinnert, 1996) as an evaluation criterion due to its equivalence to the construct of *comprehensibility* (i.e., ease of understanding). Following Kobayashi and Rinnert, our benchmark study, a 10-point Likert scale (*Poor* = 1, *Excellent* = 10) was used for the purpose of evaluation. Second, to explore the effect of text features on L2 readers' text comprehensibility judgments, we prepared short English essays as stimuli. Since Kobayashi & Rinnert suggest that rhetorical patterns, lexico-grammatical errors, and coherence are "the three frequently evaluated traits" (p. 402) in L2 academic writing, we decided to use these three constructs to differentiate the features of the texts to be evaluated. In Saito et al. (2019), the participants were asked to rate 50 speech samples that differed in terms of pronunciation, lexico-grammatical accuracy, and fluency, three major evaluation traits of L2 speech. Following that methodological model, we therefore used 16 different L2 texts that systematically differed in the focal characteristics of rhetorical patterns, lexico-grammatical errors, and coherence (i.e., disrupted sequences of ideas). In that way, we could investigate the interaction between how the readers' "ease of understanding" would interact with their perception of the three traits manifested in their scores for "clarity of meaning (RQ1) as well as the interaction between how such behavior interacted with their background (RQ2).

More specifically, to examine the three potentially influential traits of these texts, we adapted 16 different four-paragraph student essays on the topics of Cars and TV (8 about Cars and 8 about TV) taken from Kobayashi and Rinnert (1996). These essays were systematically manipulated in terms of the following characteristics: rhetorical patterns (deduction vs induction), lexico-grammatical errors (25 errors throughout the four paragraphs), and disrupted sequences of ideas (scrambled sentences in the second and third paragraphs; see Table 1 and Appendix A-2). Furthermore, we revised the "Cars" versions because these produced unexpected results in Kobayashi and Rinnert, arguably due to complications originating in specific discoursal expressions and what Kobayashi and Rinnert call the "American" (p. 405) deductive approach vs the "Japanese … general-to-specific" (p. 406) pattern. We therefore replaced the "American vs Japanese" rhetorical contrast with a "deductive vs inductive" contrast adapted from Shi and Kubota (2007). That is, when the main idea appeared in the first paragraph, we called this ordering *deductive,* but when it appeared in the fourth (last)

Table 1
Overview of Features to be Evaluated in Texts

Text number	Topic	Rhetorical pattern: Deductive (D) or Inductive (I)	Lexico-grammatical errors (+, −)	Disrupted sequences of ideas (+, −)
1	TV	D	−	−
2			+	−
3			−	+
4			+	+
5		I	−	−
6			+	−
7			−	+
8			+	+
9	Cars	D	−	−
10			+	−
11			−	+
12			+	+
13		I	−	−
14			+	−
15			−	+
16			+	+

Note. Lexico-grammatical errors categories included: word choice (4), word order (2), adverbial connectives (4), word class (2), verb form (3), number agreement (2), preposition choice (3), pronoun (1), and 4 article errors (4). Local to more global, developmental, and interlingual errors were taken from actual student errors in sets of papers on these topics (see Kobayashi & Rinnert, 1996, p. 408).

paragraph, we called it *inductive*. Finally, unlike Kobayashi and Rinnert's 16 versions, we standardized all versions to approximately 300 words, according to Shi and Kubota arguably the minimal length for an opinion essay, whether inductive or deductive, by shortening some versions to control experimental conditions and reduce participant fatigue.

We judged these different versions to be equivalent to the 50 spontaneous speech samples used in Saito et al. (2019) for the purpose of eliciting comprehensibility judgements, for two reasons. First, although the 16 versions were the result of manipulations, the originals consisted of authentic texts written by a population similar to our present participants who were often asked to evaluate L2 writing as peers, colleagues, or novice teachers (Ferris & Hedgcock, 2023; Hyland, 2019). We limited the number of essays to 16 because the results of a pilot study with 10 participants from a population similar to that used in the subsequent experiment revealed that 16 was the maximum they could focus on without undue stress (Dörnyei, 2003).

3.3. Evaluation procedure

Throughout the experiment, participating readers worked alone online but were helped by the first or second author whenever necessary. They first read about the purpose of the research and signed an informed consent form online (by clicking "Yes"), which explained that they were free to withdraw from the study during or after data collection without consequences and that their identity would not be revealed were the study to be published. They then responded to a background survey (see Section 3.5) followed by the evaluation task. Following an explanation of the 10-point comprehension rating scale and procedure, they began rating each of the 16 L2 texts at their own pace. More specifically, they were shown one of the 16 English texts one at a time and were asked to give an intuitive score for comprehensibility from 1 (*Poor*) to 10 (*Excellent*) on a Likert scale, which appeared at the bottom of the screen. The order of appearance of the 16 texts was randomized for each reader. After recording a score for each L2 text, they clicked on the *Next* icon and proceeded to the next L2 text. They were allowed to take short breaks between texts due to the relatively large number of texts to be read. Meanwhile, the first and second authors continuously monitored progress through the online platform as the participants engaged in the task.

3.4. Grouping readers

First, we grouped readers into Strict and Lenient evaluators based on the results of the above comprehensibility rating. Overall, a mean comparison[1] between the 100 Japanese readers ($M_{score} = 6.96$, $SD = 1.12$, *Range* 4.6–9.3) and 15 L1-speaking readers ($M_{score} = 6.35$, $SD = 0.52$, *Range* 5.2–6.41) suggested that the Japanese readers were slightly more lenient ($U = 412$, $p = 0.005$) than our 15 English-speaking baseline readers.

Second, the 100 Japanese readers were divided into two groups (Group 1 = 79; Group 2 = 21) via hierarchical cluster analysis using Ward's method of minimum variance with Euclidean square distance intervals (Fig. 1).

With the division generating relatively homogeneous inter-group rating behavior and high inter-rater reliability ($\alpha = 0.90$ for Group 1; $\alpha = 0.88$ for Group 2), scores were averaged to generate group mean scores (see Table 2 for the descriptive statistics) and then compared to the mean scores of the baseline L1-speaking readers' mean scores (Table 3; Fig. 2).

The results of a one-way ANOVA showed that differences in average rating scores between the three groups were statistically significant ($F\,[2, 112] = 74.217$, $p < 0.001$, $\eta p^2 = 0.57$). A post-hoc multiple comparison analysis also demonstrated that Group 1 ($M = 7.38$, $SD = 0.78$) assigned significantly more lenient scores to the 16 texts than did the L1-speaking readers ($M = 6.32$, $SD = 0.05$). The differences in rating scores between these groups showed large effect size ($d = 1.48$). In addition, the baseline L1-speaking readers assigned significantly more lenient scores than did Group 2 ($M = 5.35$, $SD = 0.66$), also with large effect size ($d = 1.92$). A visual representation of the group comparison is shown in Fig. 2. These results support the conclusion that Group 1 can be justifiably called Lenient and Group 2 Strict.

3.5. Background survey

To elicit reader variables, we created a tailor-made questionnaire modeled after the background survey in Saito et al. (2019) and designed to explore the participants' language learning experience, including types and length of English classes taken, frequency of English use in written or spoken communication, and metacognitive characteristics (i.e., self-perceived proficiency and metacognitive awareness of language learning; recall Section 2.2). Given the findings of Saito et al. and the differences between their study and ours in terms of materials and participants, we added five items (marked as * in Table 4) to the questionnaire used in Saito et al. (for the full questionnaire, see Appendix B). Table 4 shows the descriptive statistics for each variable.

3.6. Reduced categories

A set of 20 reader variables collected from the background questionnaire was reduced to thematic categories to avoid multi-collinearity and to identify any underlying constructs. To capture the background characteristics of all 100 readers, an exploratory

[1] Since a Shapiro-Wilk normality test revealed that the variable was not normally distributed, we decided to use a Mann-Whitney *U* test, which allows researchers to compare group means without assuming that values are normally distributed.

[2] Copyright obtained from John Wiley & Sons on January 7, 2023.

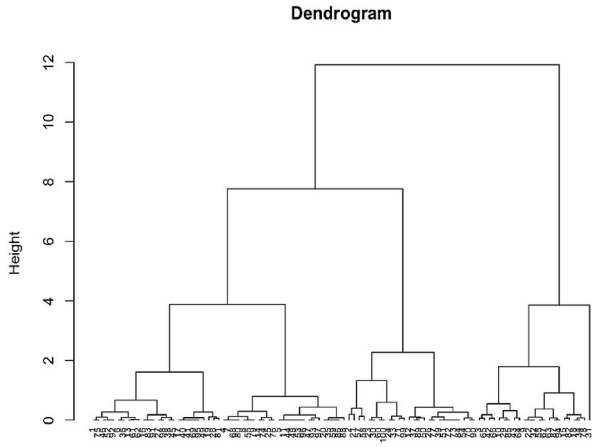

Fig. 1. Dendrogram of Three of Hierarchical Clusters based on Japanese Readers' Comprehensibility Scores

Table 2
Descriptive Statistics for L1-English vs. L2 English (Japanese) Evaluation Scores

	Group	*M*	*SD*
Comprehensibility	L1 English	6.32	0.05
	L2 English (Japanese)	6.95	1.12

Table 3
Descriptive statistics for group evaluation scores

Group	*n*	*M*	*SD*
Group 1 (Lenient Japanese)	79	7.38	0.77
Group 2 (Strict Japanese)	21	5.35	0.66
L1	15	6.32	0.05

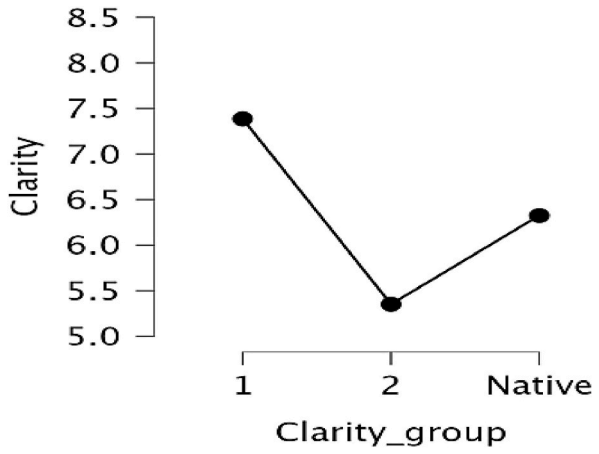

Fig. 2. Group comparison of rating scores

factor analysis was conducted, with 20 reader variables initially submitted to a Kaiser-Meyer-Olkin (KMO) test (0.62) and Bartlett' test of sphericity ($\chi^2 = 1193.78$, $p < 0.001$) and shown to be adequate for further analysis. Based on Loewen and Gonulal's (2015) field-specific recommendations, a factor analysis was then computed through the maximum likelihood method and varimax rotation using R (version 4.3.1; R Core Team, 2023). By examining eigenvalues greater than 1 and scrutinizing the scree plot, we discerned a five-factor solution as the most coherent representation of the data. This five-factor solution explained 86% of the variance, and Cronbach's alpha values for each factor were above = 0.64. To identify significant factor loadings, we adhered to a threshold of 0.6 (see

Table 4
Descriptive statistics for readers' background profiles

	Mean	*SD*	Min.	Max.
Biological age and accumulated learning experience				
Age	21.58	4.11	18	48
English teaching experience (years)	0.918	3.15	0	3
Length of stay in English-speaking countries (years)	1.04	2.33	0	13
Age of onset of English learning	9.69	3.27	0	17
Length of English learning (years)	11.03	4.47	3	37
Presence/absence of language-related experience				
Experience of receiving English academic writing instruction (Yes/No)	56%			
Experience of receiving Japanese academic writing instruction (Yes/No)	61%			
Experience of learning third language (Yes/No)	5%			
English learning experience in the classroom				
English classes taken in university (hours/week)	1.93	3.18	0	17
English writing classes taken in university (hours/week)	0.24	0.52	0	2
English use (speaking, listening, reading, and writing) outside university classes (hours/week)	1.65	4.72	0	28
Extracurricular English learning experience				
[a]Hours spent on online writing in English per week	0.55	1.93	0	14
[a]Hours spent on reading English academic/research papers per week	1.07	2.20	0	10
[a]Hours spent on writing English academic/research papers per week	1.00	2.58	0	15
[a]Hours spent on reading Japanese academic/research papers per week	1.48	3.25	0	20
[a]Hours spent on writing Japanese academic/research papers per week	1.79	3.42	0	20
L2 reader's metacognition				
Metacognition about the importance of comprehensibility in English (1 = disagree, 9 = agree)	7.54	1.43	3	9
Metacognition about the importance of native-like use of English (1 = disagree, 9 = agree)	4.49	2.09	1	9
Self-perceived proficiency				
Self-evaluation of writing comprehensibility	4.16	1.76	1	8
Self-evaluation of native-like writing proficiency	3.33	1.68	1	8

[a] These items were added to the original questionnaire used in Saito et al. (2019).

Saito et al., 2019 for a similar decision) (see Table 5 below). As a result, eight variables out of 20 were discarded due to either their low eigenvalues (given the threshold of 0.6) or to their loading on multiple factors.

As regards Factor 1, age, English teaching experience, and length of stay in English-speaking countries loaded together. Since those variables were related to the participants' experience of being exposed to the target language and having sufficient English knowledge for teaching, this factor was labeled *L2 Exposure.* Self-evaluation of native-like writing proficiency and self-evaluation of comprehensibility in English writing loaded together under Factor 2. Because those variables were all about participants' self-evaluation of different aspects of their English proficiency, the factor was labeled *Self-evaluation of proficiency.* Factor 3 was labeled *University English use* because English classes taken in university and hours spent writing English academic or research papers loaded together. Factor 4 was labeled *Length of English learning* because the variables that loaded together were about Age of onset and Length of English learning. Lastly, Factor 5 was labeled *Personal use of L2* because the two variables related to such feature (i.e., English use, including speaking, listening, reading, writing outside university classes, and Hours spent on online writing in English) were related to participants' use of English in their own time (such as online communication) and loaded together.

Table 5
Factor analysis of readers' background variables

	Factor 1 (L2 Exposure)	Factor 2 (Self-evaluation of proficiency)	Factor 3 (University English use)	Factor 4 (Length of English learning)	Factor 5 (Personal use of L2)
Age	**0.85**	0.05	0.04	−0.06	0.02
English teaching experience	**0.86**	−0.12	0.03	0.04	0.09
Length of stay in English-speaking countries	**0.60**	−0.06	0.30	0.18	0.07
Hours spent on writing English papers	**0.68**	0.10	0.42	−0.10	0.08
Self-evaluation of native-like writing proficiency	0.04	**0.91**	−0.08	−0.02	0.02
Self-evaluation of comprehensibility in English writing	−0.08	**0.84**	0.01	−0.03	−0.05
English classes taken in university	−0.10	−0.04	**0.58**	0.07	0.16
Hours spent on writing English academic/research papers per week	0.51	−0.06	**0.83**	0.06	0.14
Age of onset of English learning	0.02	0.02	−0.06	**−0.87**	−0.04
Length of English learning	0.41	0.01	0.02	**0.71**	0.10
English use (speaking, listening, reading, and writing) outside university classes	0.05	−0.08	0.22	0.06	**0.67**
Hours spent on online writing in English outside university classes	0.18	0.07	0.07	0.02	**0.98**

Note. Bolded entries represent higher values in the factor analyses.

3.7. Analysis

Turning to RQ1, we examined the relationship between comprehensibility evaluation scores and linguistic features in the texts. As explained above, using cluster analysis, we created two groups of L2 readers: Lenient readers and Strict readers. To verify the significance of differences between Lenient, Strict, and L1 readers, we conducted one-way ANOVAs.

Next, we applied a series of linear mixed-effects modeling analyses to each group of evaluators to identify the three potentially most relevant linguistic features attended to by each group. As explained in Section 3.2, these linguistic features consisted of rhetorical patterns, lexico-grammatical errors, and disrupted sequences of ideas (i.e., distorted coherence), and they were systematically incorporated in the 16 different texts based on those used in Kobayashi and Rinnert (1996).

Regarding RQ2, which focuses on reader-related features and how they might differentiate degrees of strictness in the evaluations, we examined five factors obtained from a factor analysis. Through *t*-tests, we investigated whether any of these factors demonstrated differentiation between Lenient and Strict readers.

4. Results

RQ1 What linguistic features (e.g., rhetorical patterns, lexico-grammatical errors, disrupted sequences of ideas) of opinion texts most strongly influence multilingual readers' evaluations of comprehensibility?

To explore how the two L2 reader groups (L2 Lenient, L2 Strict) differentially weighted rhetorical features, lexico-grammatical, and coherence errors in their judgments, a series of linear mixed-effects modeling analyses was conducted using the lme4 package (Bates et al., 2015) in R. The three fixed effects (rhetorical pattern, lexico-grammatical errors, and disrupted sequences of ideas) were determined by using a reference scheme, with *deductive* rhetorical pattern coded 0 and *inductive* coded 1, the distinction between *no lexico-grammatical errors* and *lexico-grammatical errors* was coded 0 and 1, respectively, and the categorization based on *no disrupted sequence* versus *disrupted sequence* was coded 0 and 1, respectively (Table 1). The random effect was the participants' ID because each participant evaluated the essays multiple times. Since the assumptions of regression analysis (i.e., homoscedasticity, normality, linearity) were met and no multicollinearity was detected through the inspection of the variance inflation factor (VIFs), we proceed to run the analyses. To test for every possible combination of text features, the following model was constructed:

Comprehensibility ~ rhetorical pattern* lexico-grammatical errors * disrupted sequences of ideas + (1|ID)

The results of the analyses are reported in Tables 6a and 6b. These reveal that while both Lenient and Strict readers attended to disrupted sequences of ideas (Estimate = 0.345, $SE = 0.053$, $t = 6.52$, $p < 0.001$ for Strict readers; Estimate = 0.708, $SE = 0.158$, $t = 4.48$, $p < 0.001$ for Lenient readers), the Strict readers also reacted negatively to lexico-grammatical errors (Estimate = 0.16, $SE = 0.044$, $t = 3.6$, $p < 0.001$ for Strict readers). However, when those variables were combined, they were not found to affect comprehensibility judgements.

RQ2 What features mark out Lenient and Strict L2 readers when evaluating comprehensibility in L2 writers' opinion texts?

Using the reduced readers' background profile (Factors 1–5), a group comparison between Lenient and Strict Japanese readers was conducted. Due to the uneven number of participants across groups (Lenient: $N = 79$; Strict: $N = 21$), a Welch's *t*-test was computed (Table 7). The results indicate that Lenient and Strict readers differed in terms of: (a) length of English learning, and (b) personal use of L2. In fact, Lenient readers (Group 1) spent longer than Strict readers on both learning and using English for purposes other than classes they took in university and for communicating with others online (Group 2).

5. Discussion

As an initial step in involving L2 readers who are neither L2 teachers nor testers in evaluating written L2 texts, our study compared

Table 6a
Results of Linear-Mixed Effect Analyses (Strict readers)

Strict readers	Estimate	*SE*	*t*	*p*
Intercept	7.39	0.087	84.4	<0.001
Rhetorical pattern	−0.061	0.045	−1.35	0.181
Lexico-grammatical errors	**0.16**	**0.044**	**3.60**	**< 0.001***
Disrupted sequences of ideas	**0.345**	**0.053**	**6.52**	**< 0.001***
Rhetorical pattern × lexico-grammatical errors	−0.010	0.040	−0.260	0.798
Rhetorical pattern × disrupted sequences of ideas	0.026	0.041	0.637	0.525
Lexico-grammatical errors × disrupted sequences of ideas	0.027	0.042	0.653	0.514
Rhetorical pattern × lexico-grammatical errors × disrupted sequences of ideas	0.001	0.041	0.021	0.983
Random effects	**Variances**	***SD***		
Participants	0.470	0.686		
Residual	2.13	1.46		
R^2 **(marginal)**	R^2 **(conditional)**			
0.05	0.23			

Table 6b
Results of Linear-Mixed Effect Analyses (Lenient readers)

Lenient readers	Estimate	*SE*	*t*	*p*
Intercept	5.35	0.145	37.	<0.001
Rhetorical pattern	−0.048	0.115	−0.413	0.683
Lexico-grammatical errors	0.030	0.088	0.338	0.736
Disrupted sequences of ideas	**0.708**	**0.158**	**4.48**	**< 0.001***
Rhetorical pattern × lexico-grammatical errors	−0.071	0.099	−0.725	0.474
Rhetorical pattern × disrupted sequences of ideas	−0.012	0.111	−0.108	0.915
Lexico-grammatical errors × disrupted sequences of ideas	0.042	0.101	0.414	0.682
Rhetorical pattern × lexico-grammatical errors × disrupted sequences of ideas	0.048	0.092	0.517	0.608
Random effects	**Variances**	***SD***		
Participants	0.239	0.489		
Residual	3.17	1.78		
R^2 **(marginal)**	R^2 **(conditional)**			
0.13	0.19			

Note. * indicates $p < 0.05$.

Table 7
Independent Samples *t*-Test

	t	*df*	*p*	Cohen's *d*
Factor 1	1.14	87.41	0.26	0.22
Factor 2	−1.65	97.40	0.10	−0.33
Factor 3	0.31	97.99	0.76	0.06
Factor 4	2.17	96.82	**0.03**	**0.43**
Factor 5	2.26	58.68	**0.03**	**0.44**

Note. Welch's *t*-test.

how L1 and L2 readers evaluated the comprehensibility of academic texts written in the classroom by L2 writers and what linguistic factors affected their ratings. The results revealed that some Japanese L1 readers were stricter than L1-English readers while others were more lenient (L2 Lenient < L1 English < L2 Strict), with differences being statistically significant. These results support the mixed results of previous studies, which showed that L2 student users were either stricter than L1 English speaker (e.g., Kobayashi, 1992) or vice versa (e.g., Hyland & Anan, 2006; Santos, 1988) depending on their background and experience. We subsequently examined the characteristics of both Lenient and Strict readers (RQ2) along with what linguistic features most affected their evaluations of comprehensibility (RQ1).

First, as reported above, in terms of the backgrounds of both types of L2 readers, Lenient readers started learning L2 English earlier in life, used English outside university longer, and communicated online more often than did Strict readers. Notably, learners' longer use of English online and outside university classes (e.g., factor loading: 0.98, Factor 5) appears to contribute chiefly to differences between the two types of readers. Follow-up analyses of mean comparisons further revealed that the Lenient readers tended to spend more time on writing online in English (Lenient readers: $M = 0.65$ hours vs. Strict readers: $M = 0.14$ hours; Cohen's $d = 0.323$). This finding makes an intriguing contrast with Saito et al.'s (2019), where lenient listeners "regularly used L2 English in professional settings" (p. 1134) compared to strict listeners. Our own lenient readers' frequent use of L2 English differed not only from Saito et al.'s participants' use of the L2 in terms of mode (speech vs writing) but also in terms of where it was used, namely outside the classroom in our study compared to at work for Saito's et al.'s lenient listeners. This may be due to our participants being students with time on their hands, especially when the COVID-19 pandemic was rampant, whereas Saito et al.'s participants were mostly working adults. However, the other characteristic of our lenient readers, namely their longer use of English speaking, listening, reading, and writing outside university classes, resonates with Saito et al.'s results, where lenient listeners regularly used L2 English in professional settings while interacting with multilingual speakers. Even though we excluded academic texts produced by professional writers, these lenient student readers may come closest to multilingual scholarly writers who disseminate their papers in their blogs and who have been reported to be more tolerant of their multilingual audiences (e.g., Luzón, 2018). In brief, exposure to non-classroom multilingual populations may be key to maintaining and even improving motivation to use L2 English and in turn may foster tolerance of comprehensibility issues in academic writing.

The most conspicuous difference between the results of our study and Saito et al.'s (2019) regarding rating leniency was that compared to their strict listeners, Saito et al.'s lenient listeners were characterized by their "higher levels of awareness of the importance of comprehensibility for communication" (p. 1134), whereas our study failed to show any effect of such metacognitive characteristics on the readers' leniency due to its low factor loading. A follow-up mean comparisons of the metacognitive awareness between Strict and Lenient readers further confirmed that these readers showed no significant differences in terms of metacognition (cf. $t[30.2] = 1.108$, $p = 0.28$ as it relates to the importance of comprehensibility in English ($t[27.8] = 0.709$, $p = 0.48$ for metacognition regarding the importance of native-like use of English). This difference in the role of metacognition may be due to modality (speech vs. writing), or it could be related to participants' profiles: (a) our participants were all students educated in the same country and who may have shared the same perceptions of L1-speaker supremacy (e.g., Mao & Crosthwaite, 2019); and (b) distance between

our participants' L1 and L2 (i.e., Japanese vs. English) was controlled in our study. This suggests that the influence of metacognition on readers' evaluation of texts should be investigated in a follow-up study with readers with varied linguistic and experiential profiles.

As regards how both L2 Lenient and L2 Strict readers reached their final scores for comprehensibility, our results reveal that disrupted sequences of ideas (i.e., coherence errors) were a common factor affecting such judgments in both groups. This may be specific to our participants, who study English in the classroom. Studies investigating tolerance in L2 text readers at professional levels show that such readers seem to focus more on lexico-grammatical errors (e.g., Flowerdew & Wang, 2016) even when targeting ELF writers (Carey, 2013), despite their lack of genre-related knowledge and publishing conventions (Mur-Dueñas, 2013). Returning to our participants' contexts, this finding first and foremost conveys a pedagogical message. For those studying the L2 in the classroom, coherence errors (i.e., disrupted sequences of ideas), which occurred only in the second and third paragraphs of our four-paragraph essays, may inhibit text comprehensibility more than lexico-grammatical errors scattered across all paragraphs. However, this also challenges us to investigate what linguistic aspects may inhibit comprehensibility outside academic genres.

A final notable finding regarding what affected the Strict and Lenient readers' comprehensibility ratings is that in addition to coherence errors, the Strict readers were also affected by lexico-grammatical errors whereas the Lenient readers were not. This observation leads to an interesting comparison with the results of Kobayashi and Rinnert (1996), our benchmark study. That study had four groups of participants (English L1-speaking EFL teachers, Japanese L2-speaking EFL teachers, and Japanese university students with and without English writing instruction) evaluate English texts. Results showed that the two teacher groups (as opposed to the two student groups) reacted more negatively to lexico-grammatical errors than to disrupted sequences of ideas. Furthermore, the students who had received L2 writing instruction gave a lower average score, even though it was still higher than the scores given by the two teacher groups for texts with lexico-grammatical errors. This suggests that these groups' evaluations may have been influenced by prior L2 writing instruction (1996). Connor-Linton (1995), published around the same time, reported a similar phenomenon, with Japanese EFL teachers focusing on "matters of accuracy" (p. 99) when rating English compositions. Thirty years later, we find that grammatical errors do not have as much impact as coherence errors on L2 text comprehensibility, especially for those who have used L2 outside the university longer and communicate online more often. Although such a finding may align with studies reporting multilingual participants' leniency when evaluating L2 texts across modalities (e.g., Kobayashi, 1992; Saito & Shintani, 2016; Saito et al., 2019), given the exploratory nature of our study, future research is necessary before we can make any generalizations.

6. Conclusion

Unlike most past studies, which mainly focused on multilingual writers' output (e.g., Canagarajah, 2020; Kubota, 2022), our study focused on their perceptions as readers, a rare objective in L2 writing studies. Though exploratory, the implications of our results are useful from both pedagogical and research perspectives, especially since we limited our scope to participants who learn the L2 in the classroom. First, as regards readers' characteristics, one crucial point to apply to instruction is the lenient readers' longer use of English outside university, especially in communicating online, compared to the strict readers. In this increasingly globalized world, fostering L2 readers who will readily understand and be tolerant of texts written by various types of L2 writers is desirable (e.g., Mauranen, 2018, 2020; Rozycki & Johnson, 2013). Furthermore, our results suggest that earlier encounters with the L2 may have long-lasting motivational effects on its continued use outside university, including its more frequent online use, which may eventually enable these learners to become tolerant evaluators of L2 texts encountered both inside (e.g., academically) and outside university (e.g., online) over time. Future synthetic studies based on our study and the findings of past L2 writing studies should investigate effects of classroom instruction combined with online communication with multilingual users in classroom settings over time for comparative purposes.

Another pedagogical implication of our study relates to the frequent past focus of L2 writing instruction on grammatical accuracy. This dimension has long been the main target of corrective feedback on L2 texts (e.g., Crosthwaite et al., 2022) as well as one indicator of L2 writing development (e.g., Manchón, 2020). Yet, what most affected our participants' comprehensibility ratings was not grammatical errors but disrupted sequences of ideas, a feature of L2 writing that will require greater attention when L2 learners are not only taught to write discourse-level texts but also at lower instructional levels. When L2 writers read compositions written by multilingual users like themselves, factors different from anticipated aspects of L2 texts may hinder their understanding, a point to be considered when planning L2 writing instruction.

Lastly, research approaching multilingual texts from readers' angle suggests another potential area connecting L2 writing and SLA research in that knowledge accumulation in speech studies can be applied to L2 writing studies. Because L2 writing research inspired by other SLA studies remains at the pioneering stage (Manchón, 2020), our study should be replicated by substituting either participants' characteristics or targeted L2 texts. This should be emphasized because our participants were particularly limited in terms of lingua-cultural diversity. Alternatively, studies could be replicated with newly-written texts (ours were written in 1996) or different types of L2 texts, including online writing (Li, 2021) or even multimodal writing (e.g., Hafner & Ho, 2020). Such studies should explore what constitutes comprehensibility for multilingual writers compared to corresponding L2 speakers. Having accumulated sufficient knowledge regarding comprehensibility of L2 texts in multilingual contexts, the norm in L2 speech research (e.g., Saito et al., 2016; Trofimovich & Isaacs, 2012; Trofimovich et al., 2020), we could then proceed in a further research direction and investigate how multilingual learners come to acquire these critical traits in natural settings for survival purposes or how we can teach them in the classroom, just as L2 speech researchers have started studying (e.g., Saito & Saito, 2017). These are some of the directions we should aim at in a world where people are increasingly expected to be functional multilingual writers and readers, especially in "multi--mediated, electronic, and collaborative environments" (Li, 2021, p. 1).

CRediT authorship contribution statement

Miyuki Sasaki: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **Yui Suzukida:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kotaro Takizawa:** Writing – review & editing, Data curation. **Kazuya Saito:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

Acknowledgments

The two texts in Appendix A (1) and (2) were adapted from "Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background" written by Hiroe Kobayashi and Carol Rinnert and published in 1996 in *Language Learning, 46*, shown there as Appendix1; Essays 1 and 4, for which the copyright was obtained from John Wiley and Sons on January 7, 2023 (Order No. 5463510023324). The content was revised with the generous help of Dr. Kobayashi and Dr. Rinnert from September to November 2022, which we very much appreciate. Needless to say, any remaining errors or flaws in the paper are ours. We would also like to thank Paul Bruthiaux for his valuable comments and suggestions. The preparation of this article was aided by JSPS KAKENHI Grant No. JP24K00096, No. JP20H01286, and Research Grant No. 2024C-065 Grant for Special Research Projects from Waseda University for the 2024 academic year.

Appendix A. [2]

1 Error-free Deductive Version – TV Topic

In our society, almost every family owns at least one TV set. TV functions to send information throughout the society; in addition, we can enjoy a variety of programs for entertainment and education. However, a TV set prevents us from communicating with family members, and this problem gets worse when families own multiple TV sets.

There has been a big change in the way families communicate. In the old days, family members enjoyed talking at the table, even after dinner was over. Children used to talk about what had happened in school or who they had played with after school. Parents, in turn, usually listened to such reports, while giving comments and advice when necessary. But now, most children try to finish dinner as fast as they can to rush to a TV set for their favorite program such as cartoons, or they try to watch it while eating their food very slowly. As a result, there is not much conversation going on between children and parents at the table.

This problem worsens when each household has more than one TV set. When children are unable to see their favorite cartoons on one TV set, they quickly turn to another one, which is usually placed in a different room in the house. Parents, too, often do exactly the same when their favorite baseball game is not on one TV set. Since family members spend less time together in front of the TV, they have less exchange of ideas about TV programs, too.

Although TV is fun to watch, it often creates a physical and psychological distance among family members. This can lead to a great loss of mutual understanding among them. Thus, many families are becoming victims of technology, allowing themselves to be controlled by their own TV sets.

2 Inductive Version with 25 Lexico-Grammatical + Coherence Errors – Cars Topic

Today, we live in fast moving society where people rush to work and to play. Although there are public transportation systems to keep up the fast pace in society such as trains and subways, many people find it more convenience to have their own vehicles. In fact, cars and bicycles are both very popular.

Cars can transport long distance quickly, so we can get done many things for either business or shopping. In the traffic jam, for example, cars can hardly run forward and it take a lot of time to get our destination. While driving cars is often not convenient. Also, it is time consuming to find a motor-pool for a car in a busy area. When riding, car can also protect us for stormy weather, so that we don't have to worry about getting wet or blow off the vehicle.

By the way, the biggest merit of bicycles is economic. Nevertheless, bicycles area not comfortable in bad weather; we get wet to the skin on a rainy day and feel cold in freezing winter. It makes us feel fitness and healthy. In addition, riding a bicycle provide enjoyable exercise. The price of bicycle is so reasonable that almost anyone can afford to buy one. It is great to expose to the wind during riding under a blue sky. Unlike cars, riding a bicycle does not use gas, and it does not cost expensive insurance and maintenance.

As we have seen, each vehicle has its own merits and demerits. For the reason, we cannot be chosen which one is better. It depends on the lifestyle we have, our daily activities, and our own personal needs.

Note: Versions (1) and (2) were adapted from Essays 1 and 4 in Appendix 1 of Kobayashi and Rinnert (1996).

Appendix B. Background Questionnaire Items

Basic Information

1. Age
2. Your mother tongue
3. Your mother's mother tongue Your father's mother tongue
4. Your major in university
5. Have you ever taught English before? If yes, how long? ___ years (e.g., 1–3 years)

Overseas Experience

6. Have you ever stayed overseas? If yes, how old were you? Why did you stay there?
7. What is your most recent TOEIC score?
8. When did you last take the test?

Past English Learning Experience in Classroom Setting

9. At what age did you start studying English?
10. How long have you been studying English?
11. Have you ever received academic writing instruction in English? Yes/No
12. Have you ever received academic writing instruction in Japanese? Yes/No

Other Language Background

13. Are you learning any language other than English? If yes, what is that language?

Current English Learning Experience

14. How many hours a week do you study English in university?
15. How many hours a week do you spend writing in English in university?
16. How many hours a week do you study English outside university (including reading, listening, speaking, and writing)?
17. How many hours a week do you spend writing in English outside university (e.g., writing to friends online)?

Experiences of Reading and Writing Reports or Academic Papers

18. Do you regularly read reports or papers in English? Yes/No. If yes: hours/week
19. Do you regularly write reports or papers in English? Yes/No. If yes: hours/week
20. Do you regularly read reports or papers in Japanese? Yes/No. If yes: hours/week
21. Do you regularly write reports or papers in Japanese? Yes/No. If yes: hours/week

Native/Non-Native English

22. Think about other people's or your own English. Do you agree or disagree with the following statement?
    - 22a It's no problem if someone speaks with some errors as long as their English is comprehensible. 9 = completely agree — 1 = completely disagree
    - 22b It's very important to use English like a native speaker without any errors. 9 = completely agree — 1 = completely disagree

Your own English

23. How native-like is your written English? 9 = very native-like — 1 = not at all native-like
24. How native-like is your spoken English? 9 = very native-like — 1 = not at all native-like
25. How comprehensible is your written English? 9 = very easy to comprehend — 1 = very difficult to comprehend
26. How comprehensible is your spoken English? 9 = very easy to comprehend — 1 = very difficult to comprehend

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Canagarajah, S. (2020). *Transnational autobiographies as translingual writing*. Routledge.
Carey, R. (2013). On the other side: Formulaic organizing chunks in spoken and written academic ELF. *Journal of English as a Lingua Franca, 2*(2), 207–228. https://doi.org/10.1515/jelf-2013-0013
Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the test of English as a foreign language*. Routledge.
Connor-Linton, J. (1995). Cross-cultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes, 14*(1), 99–115. https://doi.org/10.1111/j.1467-971X.1995.tb00343.x
Crosthwaite, P., Ningrum, S., & Lee, I. (2022). Research trends in L2 written corrective feedback: A bibliometric analysis of three decades of scopus-indexed research on L2 WCF. *Journal of Second Language Writing, 58*(1), Article 100934. https://doi.org/10.1016/j.jslw.2022.100934
Crowther, D., Holden, D., & Urada, K. (2022). Second language speech comprehensibility. *Language Teaching, 55*(4), 470–489. https://doi.org/10.1017/S0261444821000537
Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing*. Erlbaum.
Douglas Fir Group. (2016). A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal, 100*(S1), 19–47. https://doi.org/10.1111/modl.12301
Ferris, D. R., & Hedgcock, J. S. (2023). *Teaching L2 composition: Purpose, process, and practice* (4th ed.). Routledge.
Flowerdew, J. (2022). Models of English for research publication purposes. *World Englishes, 41*(4), 571–583. https://doi.org/10.1111/weng.12606
Flowerdew, J., & Wang, S. H. (2016). Author's editor revisions to manuscripts published in international journals. *Journal of Second Language Writing, 32*, 39–52. https://doi.org/10.1016/j.jslw.2016.03.004
Foote, J. A., & Trofimovich, P. (2018). Is it because of my language background? A study of language background influence on comprehensibility judgments. *Canadian Modern Language Review, 74*(2), 253–278. https://doi.org/10.3138/cmlr.2017-0011
Hafner, C. A., & Ho, W. Y. J. (2020). Assessing digital multimodal composing in second language writing: Towards a process-based model. *Journal of Second Language Writing, 47*, Article 100710. https://doi.org/10.1016/j.jslw.2020.100710
Hyland, K. (2019). *Second language writing* (2nd ed.). Cambridge University Press.
Hyland, K., & Anan, E. (2006). Teachers' perceptions of error: The effects of first language and experience. *System, 34*(4), 509–519. https://doi.org/10.1016/j.system.2006.09.001
Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehensibility scale for English-medium universities. *Language Testing, 35*(2), 193–216. https://doi.org/10.1177/0265532217703433
Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*(1), 81–112. https://doi.org/10.2307/3587370
Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning, 46*(3), 397–437. https://doi.org/10.1111/j.1467-1770.1996.tb01242.x
Kubota, R. (2022). Decolonizing second language writing: Possibilities and challenges. *Journal of Second Language Writing, 58*, Article 100946. https://doi.org/10.1016/j.jslw.2022.100946
Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing, 34*(3), 321–336. https://doi.org/10.1177/0265532216663991
Li, M. (2021). *Researching and teaching second language writing in the digital age*. Springer.
Loewen, S., & Gonulal, T. (2015). Exploratory factor analysis and principal components analysis. In *Advancing quantitative methods in second language research* (pp. 182–212). 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. https://doi.org/10.1075/jslp.3.2.01lud
Luzón, M.-J. (2018). Constructing academic identities online: Identity performance in research group blogs written by multilingual scholars. *Journal of English for Academic Purposes, 33*, 24–39. https://doi.org/10.1016/j.jeap.2018.01.004
Manchón, R. M. (Ed.). (2020). *Writing and language learning: Advancing research agendas*. Benjamins.
Mao, S. S., & Crosthwaite, P. (2019). Investigating written corrective feedback: (Mis)alignment of teachers' beliefs and practice. *Journal of Second Language Writing, 45*, 46–60. https://doi.org/10.1016/j.jslw.2019.05.004
Mauranen, A. (2018). Second language acquisition, world Englishes, and English as a lingua franca (ELF). *World Englishes, 37*(1), 106–119. https://doi.org/10.1111/weng.12306
Mauranen, A. (2020). Good texts in non-standard English: ELF and academic writing. In K. Murata (Ed.), *ELF research methods and approaches to data and analyses: Theoretical and methodological underpinnings* (pp. 71–80). Routledge.
Ministry of Education, Culture, Sports, Science and Technology, Japan. *Kyouiku Shinkou Kihon Keikaku [Fundamental Plan for the Promotion of Education]*, (2023). Retrieved from https://www.mext.go.jp/a_menu/keikaku.
Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x
Mur-Dueñas, P. (2013). Spanish scholars' research article publishing process in English-medium journals: English used as a lingua franca? *Journal of English as a Lingua Franca, 2*(2), 315–340. https://doi.org/10.1515/jelf-2013-0017
Nagle, C. (2018). Motivation, comprehensibility, and accentedness in L2 Spanish: Investigating motivation as a time-varying predictor of pronunciation development. *The Modern Language Journal, 102*(1), 199–217. https://doi.org/10.1111/modl.12461
O'Brien, M. G. (2014). L2 Learners' assessments of accentedness, fluency, and comprehensibility of native and nonnative German speech. *Language Learning, 64*(4), 715–748. https://doi.org/10.1111/lang.12082
O'Brien, M. G. (2016). Methodological choices in rating speech samples. *Studies in Second Language Acquisition, 38*(3), 587–605. https://doi.org/10.1017/S0272263115000418
Ortega, L. (2018). Ontologies of language, second language acquisition, and World Englishes. *World Englishes, 37*(1), 64–79. https://doi.org/10.1111/weng.12303
Pennycook, A. (2017). *The cultural politics of English as an international language*. Routledge.
R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org.
Rozycki, W., & Johnson, N. H. (2013). Non-canonical grammar in Best Paper award winners in engineering. *English for Specific Purposes, 32*(3), 157–169. https://doi.org/10.1016/j.esp.2013.04.002
Saito, K., & Shintani, N. (2016). Do native speakers of North American and Singapore English differentially perceive comprehensibility in second language speech? *TESOL Quarterly, 50*(2), 421–446. https://doi.org/10.1002/tesq.234
Saito, K., Tran, M., Suzukida, Y., Sun, H., Magne, V., & Ilkan, M. (2019). How do second language listeners perceive the comprehensibility of foreign-accented speech? Roles of first language profiles, second language proficiency, age, experience, familiarity, and metacognition. *Studies in Second Language Acquisition, 41*(5), 1133–1149. https://doi.org/10.1017/S0272263119000226
Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*, 217–240. https://doi.org/10.1017/S0142716414000502
Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research, 21*, 589–608. https://doi.org/10.1177/1362168816643111
Santos, T. (1988). Professors' reactions to the academic writing of nonnative-speaking students. *TESOL Quarterly, 22*(1), 69–90. https://doi.org/10.2307/3587062
Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18*(3), 303–325. https://doi.org/10.1177/026553220101800303
Shi, L., & Kubota, R. (2007). Patterns of rhetorical organization in Canadian and American language arts textbooks: An exploratory study. *English for Specific Purposes, 26*(2), 180–202. https://doi.org/10.1016/j.esp.2006.08.002
Statista. (2023). The most spoken languages worldwide. Retrieved from http://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide.

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language and Cognition, 15*(4), 905–916. https://doi.org/10.1017/S1366728912000168

Trofimovich, P., Nagle, C. L., O'Brien, M. G., Kennedy, S., Taylor Reid, K., & Strachan, L. (2020). Second language comprehensibility as a dynamic construct. *Journal of Second Language Pronunciation, 6*(3), 430–457. https://doi.org/10.1075/jslp.20003.tro

Zhang, L. J., & Cheng, X. (2021). Examining the effects of comprehensive written corrective feedback on L2 EAP students' linguistic performance: A mixed-methods study. *Journal of English for Academic Purposes, 54*, Article 101043. https://doi.org/10.1016/j.jeap.2021.101043

Zhang, X., Lu, X., & Li, W. (2022). Beyond differences: Assessing effects of shared linguistic features on L2 writing quality of two genres. *Applied Linguistics, 43*(1), 168–195. https://doi.org/10.1093/applin/amab007