Effects of Explicit Phonetic Instruction in Shaping L2 Perceptual Cue Weighting: Evidence from English /i:/-/ɪ/

Ziyue Zheng[a,*], **Elizabeth Wonnacott**[a]

[a]Department of Education, University of Oxford, Oxford, Oxfordshire, UK

[*]**Corresponding Author:**

Ziyue Zheng: ziyuezheng99@outlook.com

Tel: (+86) 181 6019 0275

Conflict of Interest: The authors have declared that no competing interests existed at the time of publication.

Abstract

Purpose: This study examined how Chinese L2 learners of English weight vowel duration and spectral cues in the English /i:/-/ɪ/ contrast at both the group and individual levels. It further explored the role of explicit phonetic instruction from correlational and causal perspectives. Finally, it investigated whether the benefits of training were uniform across learners or varied according to individual differences.

Method: The current research adopted a pre-post design with 48 Mandarin-speaking L2 English learners randomly assigned to a control group (*N* = 23) and an experimental group (*N* = 25). In the pretest and posttest, both groups completed a forced-choice identification task using synthesized 'bead-bid' continua co-varying in spectral and temporal dimensions. Between the pre and posttest, the experimental group received explicit phonetic instruction designed to raise their metalinguistic awareness of how vowel spectrum and duration functionally differentiate/i:/ and /ɪ/.

Results: At pretest, at a group level, participants primarily relied on duration, though individual variability was substantial and aligned with the perceptual learning stages posited by the Second Language Linguistic Perception model (L2LP) (Escudero, 2000, 2005). Explicit instruction reduced reliance on duration in the experimental group, while both groups increased reliance of the spectral cue. Instructional benefits varied across individuals: learners with initially more balanced cue use and those with prior immersion experience were more likely to shift away from duration.

Conclusions: These findings support the generalizability of the L2LP model to Mandarin listeners and demonstrate that explicit instruction can effectively promote cue reweighting, particularly when targeting non-contrastive dimensions in the learners' L1. Instructional outcomes were shaped by prior language experience.

Keywords: English vowels, categorical perception, explicit phonetic instruction

1 Introduction

The different sounds of a language are distinguished by acoustic properties, which combine to distinguish phonological contrasts. For example, in the case of the lax-tense English vowel contrasts, the lax vowel /ɪ/ (as in 'ship') has a shorter duration and narrower spectrum between the frequencies of the first (F1) and second (F2) formant, while its tense counterpart /i:/ (as in 'sheep') has longer vowel length and larger spectrum (Crystal & House, 1988; Hillenbrand et al., 1995). Listeners rely on these acoustic cues to distinguish phonetic contrasts; however, the multiple dimensions do not necessarily contribute equally to the acoustic differences within a contrast (Holt & Lotto, 2006; Nittrouer, 2004). L1 and L2 listeners may assign different weightings to the different dimensions (e.g., Baek, 2022; Escudero et al., 2009; Lipski et al., 2012). The current research builds on previous research investigating how English-L2 (Mandarin-L1) listeners distinguish the English /i:/-/ɪ/ contrast, and explores the role of explicit instruction in shaping L2 perception. In the following sections, we first review the literature on the perception of the /i:/-/ɪ/ contrast in native and non-native listeners, and factors affecting Mandarin-L1 listeners' cue weighting this contrast, before turning to consider the literature exploring phonetic training of this contrast in L2 listeners.

1.1 Cue weighting of the /i:/-/ɪ/ contrast in native and non-native listeners

In native perception of the English /i:/-/ɪ/, cue weighting varies across dialects. There have been abundant speech perception studies which found native American English listeners to rely strongly and sometimes exclusively on vowel spectrum (e.g., Bohn, 1995; Yu, 2023). Southern British English listeners tend to use a combination of the spectral and temporal cues, though with a stronger weighting on spectrum (Escudero & Boersma, 2004). In contrast, 16 out of 20 Scottish English listeners in the same study relied exclusively on spectral cues. In terms of L2 cue weighting in English, compared with native English listeners, extensive studies have found that, at the group level, L2 learners with various L1 backgrounds such as Japanese (e.g., Udovchenko, 2016; Yazawa et al., 2020) Spanish (e.g., Bohn, 1995; Escudero & Boersma, 2004) and Mandarin (Bohn, 1995; Wang & Munro, 1999; Yang et al., 2016; Yu, 2023) rely heavily on the vowel duration.

Moreover, research has consistently demonstrated that L2 learners differ substantially in their cue-weighting strategies, with such individual differences showing stability over time (Idemaru et al., 2012; Kong & Edwards, 2016; Schertz et al., 2016) and across contrasts (Clayards, 2018). Some studies have approached individual difference by exploring its association with learners' background variables such as language experience and L2 proficiency (e.g., Grenon et al., 2019; Ingvalson et al., 2011; Kong & Edwards, 2016), though findings remain mixed. In contrast, Escudero's Second Language Linguistic Perception model (L2LP) explains such individual variability from another standpoint, viewing them as different stages in perceptual learning (Escudero, 2000, 2005). Specifically, for Spanish-speaking learners' cue weighting of the the /i:/–/ɪ/ contrast, L2LP outlines the following stages[1]: (0) no perceptual distinction between the two vowels, with both assimilated to Spanish /i/; (1) reliance on duration (according to the L2LP model, it is easier to create a new category along the previously unused dimension (in this case, duration) than split a used dimension in L1 (in this case, vowel spectrum); (2) predominant reliance on duration with emerging sensitivity to spectral differences as exposure to English increases; and (3) eventual prioritization of spectral cues, with duration serving a supportive role.

Although the L2LP model was originally developed to explain the individual variability in Spanish learners' /i:/–/ɪ/ perception, the different learning stages it proposes should also apply to Mandarin-speaking learners. This is because, like Spanish, Mandarin possesses only a single high front vowel (/i/) and does not use vowel duration contrastively to distinguish vowels. There is some empirical support for such parallels (e.g., Wang & Munro, 1999; Zheng, 2023). For instance, Wang and Munro (1999) found that among 14 Mandarin learners, eight relied exclusively on duration, four primarily used spectral cues, and two employed both, reflecting patterns that broadly correspond to the stages outlined in L2LP. However, most previous studies on Mandarin-speaking learners have relied on

[1] Morrison (2008) proposed an intermediate 'Stage ½' between Stage 0 and Stage 1 in Escudero's L2LP model, termed 'multidimensional-category-goodness-difference assimilation' to Spanish /i/. In this stage, English vowel tokens are assimilated to Spanish /i/ based on perceived goodness: better exemplars of Spanish /i/ are labeled as English /ɪ/, while poorer ones are labeled as /i:/. Further discussion of this stage appears in Section 4.1.

descriptive statistics and subjective categorization methods. To address this, the present study followed Morrison (2008) to adopt a more systematic approach by applying hierarchical cluster analysis to identify distinct cue-weighting profiles among Mandarin listeners and examine their alignment with L2LP's proposed different learning stages. This approach not only enables a more principled account of individual differences in Mandarin listeners' cue weighting for the /i:/–/ɪ/ contrast, but also provides a means to examine whether these L2LP model's proposed stages can be generalized to languages other than Spanish.

1.2 Factors shaping Mandarin listeners' perceptual cue weighting of English /i:/-/ɪ/

One of the most influential factors shaping L2 perceptual cue weighting is the transfer of the L1 cue-weighting patterns. For instance, Japanese listeners distinguished English /i:/-/ɪ/ primarily based on duration (Grenon et al., 2019; Udovchenko, 2016; Yazawa et al., 2020), which is likely due to the L1-based assimilation (mapping English /i:/ to Japanese /ii/ and English /ɪ/ to Japanese /i/). According to the Perceptual Assimilation Model (PAM) (Best, 1995), Japanese listeners assimilate English /i:/ and /ɪ/ respectively to two distinct L1 categories /ii/ and /i/, forming a Two-Category assimilation pattern within a shared perceptual space. The Speech Learning Model (SLM) (Flege, 1995) offers a similar account: Japanese listeners' primary reliance on duration arises from equivalence classification, where L2 sounds are perceived as sufficiently similar to existing L1 categories and thus fail to trigger the formation of a new category. In this case, English /i:/ and /ɪ/ are respectively assimilated to Japanese /ii/ and /i/. Therefore, because the Japanese /ii/–/i/ contrast is primarily distinguished by duration rather than spectral cues, Japanese learners would rely heavily on duration rather than vowel spectrum when perceiving the English /i:/–/ɪ/ contrast.

However, transfer of L1 cue-weighting strategies relies on the precondition that relevant phonetic cues are either shared or differ between the L1 and L2, which might not extend to cases where L1 lacks the relevant cues altogether, as in the perception of English /i:/-/ɪ/ by Mandarin listeners. Mandarin only has one vowel category (/i/) in the high-front acoustic vowel space, whose acoustic properties are closer to the English /i:/. Critically, Mandarin does not use either vowel

duration or spectral distinctions robustly in the high-front vowel space to signal phonemic contrasts. Consequently, both the SLM and PAM would predict that Mandarin learners assimilate both English /i:/ and /ɪ/ to the single Mandarin /i/ category. Under such assumption, no clear distinction for cue preference should emerge. Yet, most learners still relied heavily on duration but not vowel spectrum when distinguishing these vowels (e.g., Bohn, 1995; Wang & Munro, 1999; Yang et al., 2016). Therefore, such preference for duration cannot be fully explained by cross-linguistic assimilation alone. Several alternative explanations have been proposed to account for L2 English learners' reliance on duration cues, particularly among speakers of languages such as Mandarin and Spanish, which have only a single vowel category in the high-front vowel space.

The first explanation is Bohn's 'Desensitization Hypothesis' (Bohn, 1995), which posits that because Mandarin only has one high-front vowel, Mandarin-L1 listeners do not use spectral cues to distinguish vowel contrasts in the high-front vowel space. Thus, they become 'desensitized' to the spectral cue in that vowel space and instead rely on duration when differentiating /i:/-/ɪ/ because vowel duration is more universally salient compared to the spectral cue, as it is more robustly encoded by the human auditory system (Bohn, 1995; Kajouj & Kager, 2019; Kivistö-de Souza et al., 2017). Because such highly discriminable cues are acoustically informative regardless of listeners' L1 background, listeners are likely to assign more perceptual weight to this cue when discriminating vowel contrasts.

The second account is proposed by the L2LP model. According to L2LP, new phonetic categories in a second language can be formed relatively quickly and successfully when the target language contrast involves an acoustic dimension that is unused in the learner's L1 (Colantoni et al., 2015). This is because constructing a new category along a previously unused dimension is cognitively easier than splitting an existing category along a dimension already employed in the L1, as the unused dimension serves as a 'blank slate' for category formation (Escudero & Boersma, 2004, p. 580). In the case of the English /i:/–/ɪ/ contrast, Mandarin has a five-vowel system /i, u, y, e, a/ (Hao, 2018), which distinguishes vowels primarily through spectral (formant-based) cues. Vowel duration, by

contrast, is not used contrastively in Mandarin to distinguish vowels. Therefore, following the L2LP framework, Mandarin listeners are predicted to rely on duration to distinguish /iː/ and /ɪ/. This offers an explanation for Mandarin listeners' predominant use of duration cues in the initial learning stages of acquiring the English /iː/–/ɪ/ contrast. The 'Desensitization Hypothesis' and L2LP's 'Blank Slate Hypothesis' differ fundamentally in how they conceptualize the role of duration in L2 speech perception. The Desensitization Hypothesis posits that duration is a deeply ingrained perceptual cue embedded in the auditory system. Consequently, it implies that learners' reliance on duration may be relatively resistant to change. In contrast, L2LP assumes that duration is not a well-established cue when it is unused in learners' L1, and thus may provide a cognitively flexible foundation for forming new phonetic categories. As a result, the two frameworks might offer divergent perspectives on the potential effectiveness of short-term perceptual training in shifting learners' attention away from vowel duration. While the Desensitization Hypothesis would suggest that such reliance may be difficult to modify, L2LP would allow for the possibility that cue weighting along an unused dimension like duration might be more susceptible to adjustment, though this remains an assumption.

The third explanation for L2 learners' over-reliance on duration is the explicit instruction they previously received. Several studies suggest that English as a Foreign Language (EFL) teachers over-emphasize vowel duration when teaching the differences between English lax-tense vowel contrasts (Flege et al., 1997; Schertz et al., 2015; Wang & Munro, 1999, 2004). However, most of these studies relied on anecdotal observations rather than directly examining the statistical relationship between prior instruction and cue-weighting patterns. For instance, Flege et al. (1997) speculated that Chinese and Korean students' heavy use of vowel duration might result from inaccurate classroom instruction. However, they did not assess exactly what kind of prior instruction participants had actually received. Wang and Munro (1999) went further and collected self-reported data from L1-Mandarin participants, all of whom reported to have been instructed that the only difference between /iː/ and /ɪ/ lies in duration. This aligns with their behavioral findings that learners

predominantly relied on the temporal cue[2]. However, there was no comparison group with accurate prior instruction. In a unpublished Master's thesis, Zheng (2023, Experiment 1) found that Mandarin speaking L2 learners who reported having received 'correct' instruction (instruction covering both spectrum and duration) of the /i:/-/ɪ/ contrast relied more on spectral cues than those with 'incorrect' instruction (instruction focusing only on duration), suggesting that prior classroom instruction experience may shape cue weighting. However, the evidence was correlational and based on self-report. The present study builds on this by using an experimental phonetic training design to test whether explicit instruction on both cue dimensions can causally shift Mandarin listeners' perception of English /i:/-/ɪ/. Before presenting our study, we first consider the more general literature on phonetic training and the evidence that we can shift cue-weighting patterns through direct intervention.

1.3 Perceptual Phonetic training to shift cue-weighting patterns

A mismatch between L2 learners' cue-weighting strategies and the acoustic properties of the target dialect may reduce comprehension accuracy. For example, Mandarin listeners tend to rely heavily on vowel duration when distinguishing English /i:/–/ɪ/, but this cue is unreliable because vowel length varies with phonetic context (e.g., vowels are longer before voiced consonants) and across dialects (Morrison, 2008). So, learners relying solely on duration are thus likely to struggle with /i:/–/ɪ/ distinctions across dialects, which may partly explain their low comprehension accuracy for this contrast (Nishi & Kewley-Port, 2007; Ruan & Saito, 2023). This raises the question of whether targeted phonetic training can reduce reliance on duration and promote greater use of spectral cues. Whether and what types of training can enhance listeners' ability to accurately identify phonetic contrasts is a key question for L2 speech perception.

One implicit method to increase reliance on less-weighted acoustic cues is exposing learners to rich input that enhances the target dimension. L2LP specifically posits that L2 learners can form

[2] Throughout this paper, the term *temporal cue* refers specifically to *vowel duration*, which is the only temporal dimension examined in this study. The terms *temporal cue* and *duration* are used interchangeably.

categories along underweighted dimensions through distributional learning, a mechanism also fundamental to infant speech category acquisition (Maye et al., 2002). In bimodal distributional training, participants are passively exposed to a series of auditory stimuli designed to target a specific speech contrast. The stimuli are arranged along an acoustic continuum (e.g., a continuum of vowel duration from short to long), but tokens near the two endpoints are presented most frequently, creating a 'bimodal' frequency distribution. This distribution of input implicitly emphasizes the existence of two distinct categories along this dimension. Research has shown that such training can improve adult L2 learners' L2 speech contrast categorization accuracy and shift their L2 cue weighting (e.g., Escudero et al., 2011; Wanrooij et al., 2013). For example, Wanrooij et al. (2013) trained Spanish-L1 Dutch learners who relied predominantly on duration to attend more to spectral differences in Dutch /ɑ/-/aː/ by presenting formant-varying stimuli with vowel duration held constant. After just two minutes of bimodal distributional learning, learners improved in vowel discrimination accuracy and increased their use of spectral cues, although duration use also intensified.

Another implicit method enhances the less-weighted cue by increasing the variability of the less-weighted cue in the training stimuli (Holt & Lotto, 2006; Kachlicka et al., 2024). This approach holds the primary cue in the training stimuli at a neutral level, making it acoustically uninformative, while varying the secondary cue along a continuum. Kachlicka et al. (2024), for instance, trained L1-Mandarin English learners to attend more to duration than pitch when perceiving prosodic contrasts including lexical stress (e.g., 'COMpound' vs. 'comPOUND') and linguistic focus (e.g., 'STUDY music' vs. 'study MUSIC'). Pitch was fixed at a midpoint while duration varied to signal contrast. After six days of training, learners significantly increased their use of duration cues. Similarly, Holt & Lotto (2006) used artificial stimuli (frequency-modulated tones defined by two acoustic dimensions, center frequency and modulation frequency) and found that passive exposure to variability in a less-weighted cue (the modulation frequency), with the dominant cue (center frequency) held constant, shifted listeners' cue-weighting patterns, which demonstrates that even brief exposure can change

perceptual strategies.

A more explicit training paradigm is High-Variability Phonetic Training (HVPT), which uses natural stimuli produced by multiple speakers across various phonetic contexts, along with immediate trial-by-trial feedback. HVPT has been shown to improve accuracy of phonetic category perception (e.g., Tremblay et al., 2023; Wiener et al., 2025). However, most studies focused on accuracy alone, and an interesting question is whether improvements in accuracy are underpinned by changes in cue weighting. One study which did explore this is the study by Zhang et al.(2021), which combined HVPT and temporally exaggerated stimuli to train L1-Mandarin English learners on English /i:/-/ɪ/. Sixty L1-Mandarin learners of English were divided into multiple-talker, single-talker, and control groups, who participated in 7 training sessions with temporally exaggerated stimuli. Results showed both multiple-talker and single-talker groups significantly improved perception and shifted reliance from temporal to spectral cues, with no group differences.

The studies reviewed above suggest perceptual training can shift L2 learners' usage of different acoustic dimensions either by implicit input exposure or more explicit feedback-based approaches. According to the Attention-to-Dimension (A2D) model, perceptual learning occurs when selective attention is redistributed—specifically, when attention is directed toward previously underweighted dimensions or away from irrelevant but dominant ones (Francis et al., 2008; Francis & Nusbaum, 2002; Kondaurova & Francis, 2010). This framework underlies the rationale for these training procedures—redirecting listeners' attention (through stimulus design or feedback) toward more informative, yet underutilized, cues and away from irrelevant but dominant cues (Conover, 2023; Francis & Nusbaum, 2002; Holt et al., 2018). In light of this, both distributional learning and cue variability training shift attention toward the less-weighted dimension rather *implicitly*, but they do not necessarily reduce attention to the less relevant, over-relied cue. As a result, listeners may begin to use the secondary dimension while still reinforcing the primary one, as seen in Wanrooij et al. (2013). HVPT, though more explicit in using feedback, still does not directly raise awareness of acoustic dimensions. Learning is largely error-driven. Given this, the A2D model raises an interesting

and important question of whether *explicit* instruction, that directly enhances learners' metalinguistic awareness how specific cues function, can direct attention to a reliable but underused acoustic cue and away from misleading ones, and thereby 'bypass' the need for an implicit training process (Andringa, 2020, p. 6)? Moreover, explicit instruction is often the most practical form of phonetic learning in classroom settings, where opportunities for rich, naturalistic/synthetic exposure to target contrasts are limited. Unlike passive listening or HVPT tasks, explicit contrast instruction remains the more common and accessible method in formal L2 education. Therefore, it is meaningful to examine whether and how explicit phonetic instruction benefits L2 listeners.

While several studies have shown that explicit instruction can improve L2 learners' accuracy in identifying contrasts (Conover, 2023; Guion & Pederson, 2007; Icardo Isasa, 2020; Ruan & Saito, 2023), only a smaller number of studies have addressed the impact of explicit instruction on cue weighting more directly. For example, Chandrasekaran et al. (2016) found that English-L1 listeners (without any prior knowledge of tonal language) who were trained to focus on pitch direction (a secondary cue for Mandarin tones) showed increased reliance on this cue. However, the 'instruction' was minimal, consisting of a single sentence suggesting focusing on pitch direction is a useful strategy, with no explanation of the acoustic cues or their functional role. For instance, participants from the 'pay attention to pitch direction' group were merely shown the instruction saying 'Prior participants have found that creating rules related to changes in pitch direction led to good task performance' (p. 572). Similarly, Jasmin et al. (2021) attempted to direct attention to pitch or loudness through brief prompts, but Mandarin listeners continued to prioritize pitch regardless of the task. In both cases, instruction did not involve metalinguistic explanation or detailed cue guidance. It is possible that the effectiveness of explicit instruction in shifting cue weighting depends on how metalinguistically richly the target cues are taught. The current research therefore investigates whether more comprehensive and metalinguistically rich instruction, that explicitly focuses on the role of specific phonetic cues, would be more effective in shifting cue-weighting patterns.

It is worth noting that one explicit training approach involves *visually* raising learners' metalinguistic awareness of the target dimension, often used in training lexical tones or prosodic features such as English stress. These methods employ visual analogs of pitch contours, such as displaying pitch trajectories (e.g., H. C. Chen & Tian, 2020; Liu et al., 2011) or using gestures including hand movements (e.g., Zhen et al., 2019) and head or neck motion (e.g., T. H. Chen & Massaro, 2008). Such visual-based interventions have been shown to improve tone and stress perception. However, these techniques may be better suited to suprasegmental features like tone and stress, where pitch direction and contour can be easily represented visually. In contrast, vowel spectrum is a more abstract acoustic property that cannot be directly visualized in the same intuitive way.

Another understudied aspect of perceptual training is whether its benefits vary across individuals. Specifically, does training alter cue weighting uniformly, or do its effects depend on learner-specific factors? Existing studies suggest variability in how individuals adjust their cue weighting after training (e.g., Grenon et al., 2019; Wanrooij et al., 2013). For instance, Grenon et al. (2019) trained Japanese listeners using a combination of HVPT and explicit phonetic instruction within a self-adaptive paradigm. Before training, participants primarily relied on vowel duration. After training, however, some shifted to spectral dominance while others remained duration-dominant. Despite these findings, the factors predicting such divergent outcomes remain unclear. In contrast, Wanrooij et al. (2013) demonstrated systematic individual differences in how L2 learners benefited from distributional learning. Their results revealed that learners with greater initial sensitivity to underused cues at pretest showed more improvement and incorporated these cues more effectively into their weighting strategies in the posttest. The present study contributes to this line of research by examining whether there are individual differences the effect of training, and which factors (such as background variables and initial cue-weighting profiles) predict these differences.

1.4 The current study

This study investigates how Mandarin-speaking learners of English weight acoustic cues in

perceiving the English /i:/-/ɪ/ contrast by examining both group-level patterns and individual variability. To systematically account for individual differences, we employ hierarchical cluster analysis to identify distinct cue-weighting profiles among Mandarin listeners. This provides a principled analysis of individual variation and allows us to assess whether these patterns align with the different learning stages proposed by the L2LP model, thus testing its generalizability beyond Spanish-speaking learners.

Based on previous findings for Mandarin learners (Bohn, 1995; Wang & Munro, 1999; Yang et al., 2016; Yu, 2023), we expected vowel duration to serve as the primary cue at the group level. According to the L2LP model, learners progress from reliance on duration to increased sensitivity to spectral cues as their exposure and proficiency grow. Therefore, at the individual level, we predicted three clusters reflecting these learning stages: 1) predominant reliance on duration, 2) mixed but transitional use of both cues (with a dominant but *reduced* use of duration and emerging spectral sensitivity), and 3) predominant reliance on spectral cues. Because higher proficiency and greater immersion might be associated with more advanced stages of speech acquisition, we also predicted that learners with higher proficiency or immersion experience would rely more on spectral cues and less on duration, compared with learners with lower proficiency and no immersion experience.

Further, we ask whether explicit phonetic instruction, which is metalinguistic in nature and directly explains the functional role of acoustic cues (the greater reliability of spectral cues over vowel duration for the English /i:/-/ɪ/ vowel contrast) can modify L1-Mandarin, L2-English learners' cue-weighting strategies. Therefore, the current study adopted an articulatory-focused training paradigm, which may more effectively guide learners' attention to spectral cues by linking them to concrete articulatory configurations. The effectiveness of this instruction will be tested using a pre-post design to test learners' weightings of the duration and spectral cues before and after explicit phonetic instruction, and comparing against a control group who did not receive the instruction. We also explore individual variability in training outcomes and the factors underlying these differences.

From a group-level perspective, we predicted that explicit training would decrease reliance on

duration and increase reliance on spectral cues, since, according to the A2D model, such instruction heightens metalinguistic awareness and directs attention toward the more reliable cue (vowel spectrum) and away from unreliable one (vowel duration). However, the L2LP model's 'blank slate' hypothesis suggests that Mandarin learners may more easily adjust along the unused duration dimension than along the already-used spectral dimension. Thus, training effects might be more pronounced for the duration cue than for spectral cue. At the individual level, we predicted variability in training outcomes. Because L2LP conceptualizes cue re-weighting as a trade-off between cues, learners who progress to a higher stage after training should show both increased spectral weighting and decreased duration weighting. Moreover, training effects may be strongest for intermediate learners, who are at a learning stage where their perceptual system is sufficiently established to benefit from explicit instruction but still flexible enough to adjust cue weighting.

Finally, we include a questionnaire on participants' prior instruction concerning this contrast to examine whether those who received inaccurate instruction would rely more heavily on duration, as observed in Zheng (2023).

2 Method

2.1 Participants

A total of 48 L1-Mandarin English learners were recruited from the authors' personal network. The study was advertised to university students from China via email and social media platforms including WeChat and WhatsApp. The recruitment advertisement specified that 1) participants should not be native speakers of any language other than Mandarin; 2) they should be L2 speakers of English at *intermediate* or *advanced* level; and 3) they should not speak any languages other than English and Mandarin at an *advanced* level. To quantify these criteria, participants rated their proficiency in each language they spoke across listening, speaking, reading, and writing respectively on a 7-point Likert scale[3]. The total score (range = 4 to 28) was divided into three equal intervals:

[3] Please refer to Section 2.3.2 for more detailed explanation about the Likert scale.

‘low’ proficiency corresponded to scores of 4 to 12; ‘intermediate’ corresponded to 13 to 20, and ‘advanced’ corresponded to 21 to 28. Summary statistics for the four skill-specific and overall proficiency scores are presented in Table 1. Based on these self-reported scores, all participants were classified as having either overall intermediate or advanced English proficiency, and none reported advanced proficiency in any third language.

[Insert Table 1 around here]

The second recruitment criterion on English proficiency was to guarantee that participants had the knowledge that the letter combinations ‘ea’ (as in ‘bead’), which would appear in the experiment, corresponded to the /i:/ sound, while ‘i’ (as in ‘bid’) corresponded to the /ɪ/ sound. To further ensure accurate label interpretation and data validity, a visual inspection of each participant’s response heatmaps was conducted. Data from four participants who exhibited reversed cue-weighting patterns—suggesting confusion about vowel-label mappings—were excluded from analysis. After this trimming, 20 participants remained in the control group and 24 in the experimental group. The remaining 44 participants were randomly assigned to the experimental (*N* = 24; 17 females, 7 males) and control groups (*N* = 20; 10 females, 10 males). The mean age of participants was 22.75 years (*SD* = 0.90, range = 21–24) in the experimental group and 22.85 years (*SD* = 2.01, range = 19–30) in the control group. The sample size per group is consistent with previous studies employing pretest–posttest training paradigms in perceptual cue weighting research (e.g., Chandrasekaran et al., 2016; Kondaurova & Francis, 2010; Wiener et al., 2025; Zhang et al., 2021).

All participants provided informed consent before completing the study. Each participant received a compensation of 20 yuan for their participation. Approval of the research was granted by the Central University Research Ethics Committee (CUREC) of the University of Oxford (Research ethics reference: EDUC_C1A_23_027).

2.2 Stimuli

Stimuli for the word recognition task in both the pre-/posttest consisted of a set of monosyllabic word pair continua: ‘bead-bid’. The original recordings used to create the stimuli were from

Giannakopoulou et al. (2017), and the recordings were natural 'bead', 'bid' speech exemplars recorded by a male native speaker of British English. The bead token had a first formant (F1) of 325 Hz, a second formant (F2) of 2510 Hz, and a duration of 287 ms, while the bid token had an F1 of 495 Hz, an F2 of 2057 Hz, and a duration of 127 ms.

2.2.1 Distribution of vowels in the target stimuli

Vowels of the target stimuli were 35 (7×5) synthetic vowels manipulated from the natural recordings, which varied in vowel duration and spectrum. The spectral dimension varied in seven steps and the temporal dimension varied in five steps. Values for the spectrum and duration of these vowels are shown in Figure 1. The horizontal and vertical axis show how duration and spectrum of the vowels vary respectively. To approximate the naturally produced speech samples, F1 and F2 of the most /i:/-like vowel (Token 5 in Figure 1) were set at 340 Hz and 2600 Hz respectively, while F1 and F2 of the most /i:/-like vowel (Token 31) were set at 488 Hz and 2114 Hz. The duration of the former was 260ms, and duration of the latter was set at 100ms. In order to form a continuum from /i:/ to /ɪ/, five linearly equal steps were created, with duration decreasing from 260ms to 100ms. At the same time, F1 and F2 co-varied in seven equal mel steps[4], where F1 increased from 340Hz to 430 Hz in steps of 25 mels and F2 decreased from 2600 Hz to 2114 Hz in steps of 30 mels.

[Insert Figure 1 around here]

2.2.2 Manipulation of stimuli

To begin with, the /i:/ vowel was cut from the natural 'bead' sample. F1 and F2 of the vowel were manipulated using the Klatt synthesizer (Klatt & Klatt, 1990) in Praat (Boersma & Weenink, 2023), yielding 7 manipulated vowels with differing F1 and F2 values: F1 (340Hz)/F2 (2600Hz), F1 (363Hz)/F2 (2514Hz), F1 (387Hz)/F2 (2430Hz), F1 (410Hz)/F2 (2349Hz), F1 (436Hz)/F2 (2268Hz), F1 (462Hz)/F2 (2190Hz), F1 (488Hz)/F2 (2114Hz). After spectral manipulation, the duration of the seven

[4] The value of power spectrum is usually converted from linear frequency scale (Hz) to the mel-scale because humans perceive sound frequency in a non-linear way. Linearly equal F1-F2 steps might not be perceptually equal to listeners. Therefore, F1 and F2 of the stimuli varied in equal mel steps rather than linear steps. The formula converting the linear frequency into mel-frequency is as follows:

$$m = 2595\, lg\left(1 + \frac{f}{700}\right)$$

vowels was manipulated via Praat's 'To Manipulation' function. This yielded five stimuli with varying duration for each of these seven vowels: 100ms, 140ms, 180ms, 220ms, 260ms. Therefore, a total of 35 synthetic vowels presented in Figure 1 were generated. Finally, vowels in the original 'bead' sample were cut and replaced with the 35 manipulated vowels. In this way, 35 'bead-bid' word stimuli were created.

2.3 Procedures

The experiment was conducted through the ''Gorilla Experiment Builder' (https://gorilla.sc/) platform (Anwyl-Irvine et al., 2020). Participants accessed the experiment online using their own computers. Task instructions appearing on the screen were given to the L2 learners received in Chinese, their native language. Before the experiment began, participants were asked to wear headphones and they completed a sound check, in which they listened to two Mandarin words trials ('北京' [Beijing] and '朋友' [Friend]) spoken by a female native Mandarin speaker. After hearing each word, they were asked to type what they heard. Participants were considered to have passed the sound check if they correctly typed both trials. If a participant failed to enter the correct words, they repeated the process. This cycle continued until they successfully typed both words correctly. Participants could not proceed to the main experiment until they passed this sound check. All participants passed the sound check. After the sound check, they gave informed consent.

The experiment adopted a pre- and posttest design and contained two sessions: In Session 1, participants completed the pretest, followed by the language background questionnaire; in Session 2, they completed an explicit training session followed immediately by the posttest. The forced-choice word recognition task used at pretest and the post test was identical for the two conditions. Session 1 took approximately 15 minutes to complete, and Session 2 took 15 minutes for the control group and 25–30 minutes for the experimental group. Participants were invited back to Session 2 no less than one day but no more than five days after the first session.

2.3.1 Pretest

This was a 2 forced-choice word recognition task. In this task, participants identified what they heard by clicking one of the two boxes presented on their screen. Each box contained an orthographic option (bid versus bead). The trial started with an 800ms presentation of a fixation (+) placed at the center of the screen. After that, there was a 100ms pause, during which participants saw a blank screen. Then, the auditory stimulus was played, and the two options were presented at the center of the screen at the same time the audio started. Participants were expected to respond after the audio had finished by clicking one of these two options, and the next trial started 100ms later. There were five blocks of the word labeling task. Each block contained the 35 'bead-bid' stimuli. In total, there were 175 trials (35 stimuli × 5 blocks). For each participant, presentations of the trials within each block were randomized. Each trial was played only once per block. Participants took a short break after completing each block.

2.3.2 Language experience questionnaire

This was presented in Mandarin. Demographic information was collected on age and gender, and participants were asked to list all the languages they spoke and rate those languages in terms of self-perceived proficiency in speaking, listening, writing and reading with a 7-point Likert scale, where 1 corresponded to 'very poor' and 7 indicated 'native like'; the next part asked participants to list the English-speaking countries or regions where they had previously stayed for over 6 months, the age they arrived at these places and how long they stayed.

2.3.3 The instruction question

We further asked participants whether they have been explicitly instructed about the differences between English /i:/ and /ɪ/. This question was not included in the language questionnaire since we were concerned that this might prompt some participants to seek out such instruction in the interval between the sessions) but instead, presented at the beginning of Session 2. Those who selected 'Yes' were asked to describe the instruction they had received Responses were coded into two categories: 'correct instruction' and 'incorrect instruction'. A response was coded as 'correct instruction' if it described differences between the mouth shape or position of the tongue

for the two vowels. Responses that did not mention these articulatory features, including but not limited to those that identified duration as the only distinguishing feature, were coded as incorrect. For instance, responses indicating that the difference only lies in duration were coded as incorrect instruction.

Among the 20 participants in the control group, two reported that they had never received instruction on the differences between /i:/ and /ɪ/; nine were coded as having received accurate instruction from the teachers; and nine were coded as having received inaccurate prior instruction. Among the 24 participants in the experimental group, one participant had never received instruction regarding the differences between the English /i:/-/ɪ/; thirteen were coded as having received inaccurate instruction; and ten were coded as having received inaccurate instruction.

2.3.4 Treatment

Participants in the experimental group were presented with materials providing detailed and explicit instruction on the differences between the two English vowels /i:/ and /ɪ/. Before studying the instruction materials, participants were told that they would learn something about the differences between the English vowels /i:/ and /ɪ/, after which they would complete a word recognition task. Then, they were presented with the learning materials on their computer screen. In the first section of the instruction materials, participants learned that the vowels differ in two aspects—vowel duration and the different shapes of the mouth when producing these sounds. The second part of the learning materials focused on exactly how the mouth shapes of the two vowels differed, and participants were told that the degree of mouth opening of /ɪ/ falls between /i:/ and /e/ and for this reason the property of /ɪ/ also falls between /i:/ and /e/. This was followed by a 2-minute video. In the first part of this video, a female native speaker of standard southern British English produced minimal pairs of /i:/ and /ɪ/, asking the audience to try to hear the difference; in the second part of the video, the same speaker produced minimal pairs of /ɪ/ and /e/. After watching the video, participants were told that duration is not so reliable since it is more affected by individual difference of the speakers. For the control group, the treatment was to watch the same video as that

in the instruction materials for the experimental group without receiving any form of instruction. The posttest began immediately after the treatment session. The forced-choice word recognition task used in the post test was identical with the one used in the pretest. For the control group, Session 2 took about 15 minutes to complete; for the experimental group, Session 2 took about 25-30 minutes.

2.4 Analysis

The analyses aimed to quantify how listeners weighted spectral and durational cues in perceiving the /iː/–/ɪ/ contrast and how these weightings changed following training. Analyses were conducted at both the group and individual levels to capture general patterns as well as individual variability in baseline cue use and training gains. All analyses were conducted in R (R Core Team, 2024). Analyses in the Results section were organized into the following structure:

1. Group-level analyses (Section 3.1)

Mixed-effects logistic regression analysis was used to quantify the extent to which a group of listeners relied on the spectral cue and duration cue when perceiving /iː/ versus /ɪ/. This approach accounts for the fact that observations from individual participants are not independent (and perceptual cue weighting is well-known for its individual variability e.g., Kong & Kang, 2023; Schertz et al., 2015). Several mixed-effects models were constructed using the lme4 package in R (Bates et al., 2015); the details of each model are reported in the relevant subsections of Section 3. Group differences were compared based on the beta value estimates (β) of fixed effects of the mixed-effects models. The basic model can be expressed using the following formula:

$$Response = \beta_{spec} \times Spec + \beta_{dur} \times Dur + a_{participant} + \varepsilon \quad (1)$$

Participants' response was coded as 1 where they picked 'bead' and 0 where they picked 'bid'. The variable Spec stands for the seven spectral steps of the acoustic stimuli. It was coded as a numeric variable centered around 0 (3, 2, 1, 0, -1, -2, -3). Likewise, the variable Dur stands for the five temporal steps of the stimuli, which were encoded as 2, 1, 0, -1, -2. Using mixed-effects logistic regression models, we examined:

1) Baseline cue-weighting patterns (Section 3.1.1):

Mixed-effects logistic regression models were used to assess how the participants as a whole relied on spectral and durational cues in the pretest. An additional model examined whether background factors—Prior Instruction (Instruction), Immersion Experience (Experience), Listening Proficiency (LProf) and Overall Proficiency (OProf)—influenced baseline cue weighting. In the model, Instruction was coded as 0.5 for participants who received accurate instruction—i.e., responses that mentioned articulatory differences such as mouth shape or tongue position—and as -0.5 for participants who received inaccurate instruction, defined as responses that did not mention articulatory differences, including those that described duration as the only cue. Experience was coded as 0.5 for participants who had lived in an English-speaking country for six months or more and –0.5 for those without such experience. Listening proficiency (LProf) was coded as a centered 7-point scale ranging from –3 to 3. Overall proficiency (OProf) was calculated by summing participants' self-ratings in reading, listening, speaking, and writing, yielding a continuous score from 4 to 28, which was also centered around 0.

2) Training's effects (Section 3.1.2):

To evaluate the intervention's effect, first a two-way mixed ANOVA was used to assess changes in response bias across test sessions and groups, followed by a mixed-effects logistic regression examining the interaction between Group and Test on cue weighting. The variable Test refers to the test condition of the observations, which had two values: the pretest condition (encoded as -0.5) and the posttest condition (encoded as 0.5). The variable labeled as Group refers to the control group (encoded as -0.5) and experimental group (encoded as 0.5).

2. Individual-level analyses (Section 3.2)

Individual differences were investigated based on the β coefficients for each participant, which were extracted from the pretest or posttest mixed-effects logistic regression models through the *Coef* function in R. Using the relevant β coefficients, we can compute cue weighting as follows (Yazawa et al., 2020):

$$Cue\ weighting = \frac{|\beta_{spec}|}{¿\beta_{spec}\vee + ¿\beta_{dur}\vee ¿¿} \quad (2)$$

Values above 0.5 indicate great reliance on the spectral cue and values below 0.5 greater reliance on the duration cue. Based on the relevant β coefficients and cue-weighting scores, we examined:

1) Baseline individual variability (Section 3.2.1):

Regression coefficients for β_Spec and β_Dur were extracted from pretest models and individual cue-weighting scores were computed. Hierarchical cluster analysis was then used to identify distinct baseline cue-weighting profiles among participants. Principal component analysis was conducted to validate the clustering structure and to explore the underlying dimensions of cue use (overall magnitude of cue use and trade-off between two cues) in each cluster identified by the hierarchical cluster analysis. Participants' pretest cue profiles were further examined in relation to background factors using Fisher's Exact Tests for categorical variables (Instruction, Experience) and Kruskal–Wallis tests for continuous variables (LProf and OProf).

2) Individual differences in cue-weighting change patterns (Section 3.2.2):

Change scores (Δβ_Spec, Δβ_Dur, Δweighting score) were calculated by subtracting pretest from posttest coefficients. Hierarchical cluster analysis was then applied to these change scores to identify profiles of cue-weighting change. Multiple linear regression analyses further tested whether these change profiles were predicted by participants' background factors (Instruction, Experience, LProf and OProf) and baseline cue-use patterns (pretest β_Spec, β_Dur and weighting scores).

3 Results

3.1 Group-level patterns

3.1.1 Baseline group-level cue-weighting pattern in the pretest

A. Baseline cue-weighting patterns

To investigate participants' (N = 44) baseline cue-weighting behavior prior to the intervention, we analyzed their pretest responses using a mixed-effects logistic regression. The formula of the model is as follows:

Response ~ Spec + Dur + (Spec + Dur | ParticipantID)

Mixed-effects model results for L2 learners' cue weighting in the pretest indicated significant positive effects of *Spec* (β = 0.45, *SE* = 0.10, z = 4.48, p < .001) and *Dur* (β = 1.32, *SE* = 0.15, z = 8.73, p < .001), with a significant intercept (β = 0.71, *SE* = 0.12, z = 5.78, p < .001). This suggests that at group level, participants relied on both the spectral and temporal cues in the pretest, but they assigned more weight to the temporal cue than the spectral cue, as estimate of the former is larger than the latter. The significant intercept indicates that participants were biased toward selecting 'bead' at baseline.

B. Influence of background factors on Baseline cue-weighting patterns

To investigate whether these baseline cue-weighting patterns were modulated by individual background factors, we fit an additional mixed-effects logistic regression model. The predictors included four factors derived from the language background questionnaire: prior instruction about the /i:/–/ɪ/ contrast (Instruction), experience of living in an English-speaking country (Experience), self-rated English listening proficiency (LProf), and self-rated overall English proficiency (OProf). Three participants who reported never receiving any instruction on the /i:/–/ɪ/ contrast were excluded from this analysis. The model included all interactions between these background variables and the two cue dimensions. The formula of the full model was:

Response ~ Spec × Experience + Dur × Experience + Spec × Instruction + Dur × Instruction + Spec × OProf + Dur × OProf + Spec × Lprof + Dur × Lprof + (Spec + Dur | ParticipantID)

Model results (see Table 2) revealed no significant (or near significant) interactions between cue dimensions and any background factor. These findings thus found no evidence that individual differences in instruction, proficiency, and experience influenced cue-weighting patterns at pretest.

[Insert Table 2 around here]

3.1.2 Cue-weighting shifts at group level

A. Intervention's effect on response bias[5]

To evaluate whether training affected response bias, we analyzed the proportion of 'bead' responses at each test session. A two-way mixed ANOVA with Group (Control/Experimental) as a between-subjects factor and Test (Pretest/Posttest) as a within-subjects factor revealed a significant main effect of Group, $F(1, 42) = 4.78$, $p = .034$, generalized $\eta^2 = .078$. However, there was no main effect of Test, $F(1, 42) = 1.29$, $p = .26$, and no significant Group × Test interaction, $F(1, 42) = 0.004$, $p = .95$. To determine whether participants displayed a general response bias, we compared the mean proportion of 'bead' responses against a chance level of 0.5. At pretest, both the control group ($M = 0.55$, $SD = 0.07$) and the experimental group ($M = 0.60$, $SD = 0.10$) showed a significant bias toward 'bead' responses, $t(19) = 2.96$, $p = .032$ and $t(23) = 4.86$, $p < .001$. At posttest, the experimental group continued to show a significant bias ($M = 0.58$, $SD = 0.10$), $t(23) = 4.18$, $p = .001$, while the control group's mean portion did not differ from chance ($M = 0.53$, $SD = 0.10$), $t(19) = 1.38$, $p = .73$. Taken together, these results suggest that participants in both groups tended to favor 'bead' responses overall, with the experimental group showing a stronger bias. However, because neither the main effect of Test nor the Group × Test interaction was significant, there is no clear evidence that training altered the bias pattern between or within groups.

B. Intervention's effect on group-level cue weighting

Response patterns for the experimental group and control group across the pretest and posttest are visualized in Figure 2[6]. To examine the interaction between test sessions and group, a mixed-

[5] Following an anonymous reviewer's suggestion, a supplementary mixed-effects logistic regression was conducted to examine response bias (Response ~ Group × Test + (Test | ParticipantID)). The results confirmed the ANOVA and descriptive statistics reported in the main text: both groups showed a baseline bias toward 'bead' responses at pretest, with the experimental group showing a slightly higher bias, and no reliable change from pretest to posttest. Full statistics are reported in Supplementary Materials.

[6] The vertical axis of the heat map presents the 7 spectral steps and the horizontal axis shows the 5 duration steps. Therefore, the 35 (5 × 7) cells in the heat graph correspond to the 35 stimuli. The color of each cell was determined by the percentage of 'bead' responses. The darker color indicates a higher percentage of the 'bead' responses, whereas lighter color means a dominance of the 'bid' responses. Therefore, reliance on the spectral cue is visualized as darker color along the upper edge of the rectangle compared with the lower edge; reliance on duration is indicated by the darker color along the right edge in comparison with the left edge.

effect logistic regression model was conducted over the pre- and posttest data from both the experiment group and control group, where the fixed effects were Spec, Dur, Group and Test. The Formula of the model is as follows:

Response ~ Spec × Group × Test + Dur × Group × Test + (Spec × Test + Dur × Test | ParticipantID)

The random effects were intercepts for each participant and by participant slopes for Spec × Test + Dur × Test, reflecting that participants may differ in the pre- and posttest in their use of the cues in generating responses. Table 3 summarizes the results of the model.

[Insert Figure 2 around here]

[Insert Table 3 around here]

The positive estimate for Spec (β = 0.69, $p < .001$) indicates that, as vowel spectrum values increased, listeners (across both groups and test sessions) were more likely to identify the stimulus as 'bead.' Similarly, the positive estimate for Dur (β = 1.17, $p < .001$) shows that participants tended to label stimuli with longer vowel duration as 'bead'. The effects of both the predictors Spec and Dur are statistically significant, providing evidence that participants overall relied on both vowel spectrum and duration when identifying the stimuli. Moreover, the fact that estimate of Dur is larger than that of Spec means the response pattern was more likely to change with the variation of duration compared with spectrum. This indicates that duration functioned as a more dominant cue for the two groups of participants overall.

Significant two-way interactions were observed for both *Spec* × *Test* (β = 0.51, $p < .001$) and Test × Dur (β = –0.25, $p = .017$). The positive estimate for *Spec* × *Test* indicates that, on average, both groups (experimental and control) increased their use of the spectral cue from pretest to posttest. The negative estimate for Test × Dur suggests that both groups decreased their reliance on the duration cue over time. Critically, the three-way interactions Spec × Group × Test and Group × Test × Dur assess whether and how cue use changed differently across groups. The Spec × Group × Test interaction was not significant (β = 0.06, $p = .772$), indicating no evidence that changes in spectral cue reliance differed between groups. In contrast, the Group × Test × Dur interaction was significant

(β = –0.60, *p* = .003), showing that the experimental group exhibited a significantly greater decrease in reliance on duration cues than the control group.

To further break down the interactions, two separate regression models were run for the control and experimental groups using the formula:

Response ~ Spec × Test + Dur × Test + (Spec × Test + Dur × Test | ParticipantID)

Table 4 summarizes the results for both groups. Consistent with the three-way interactions, both the experimental group (β = 0.53, *p* < .001) and control group (β = 0.47, *p* = .01) showed a significant positive interaction between Spec and Test. However, only the experimental group showed a significant negative interaction between Dur and Test (β = -0.61, *p* < .001), indicating a decreased reliance on vowel duration after explicit training.[7]

[Insert Table 4 around here]

3.2 Individual-level patterns

3.2.1 Baseline Individual-level features at pretest

Regression coefficients of Spec (β_Spec) and Dur (β_Dur) for each participant in pretest were extracted from the random effects of separate models of 1) the control group in the pretest and 2) the experimental group in the pretest through the *Coef* function in R. Then, a weighting score was calculated for each participant using Formula (2) in Section 2.4. Figure 3 visualizes the distribution of β_Spec, β_Dur and weighting scores for the two groups in pretest.

[Insert Figure 3 around here]

To examine whether participants demonstrated distinct patterns in their reliance on spectral and temporal cues prior to training, a hierarchical cluster analysis was conducted on their pretest cue-

[7] As suggested by an anonymous reviewer, we also assessed whether the reduced reliance on duration cues in the experimental group was associated with decreased response consistency. We conducted complementary analyses using Shannon entropy. First, Shannon entropy was calculated for each participant at each test session to assess overall certainty of the responses. Entropy values near 0 indicate highly certainty (e.g., consistently selecting 'bead' or 'bid'), whereas values near 1 reflect maximum uncertainty, with responses distributed equally across options. A mixed-effects linear regression model was fitted using the formula: entropy ~ Group × Test + (1 | ParticipantID). The results revealed no significant interaction between Group and Test (β = 0.02, SE = 0.02, t(42) = 0.77, *p* = .45), nor a significant main effect of Test (β = -0.01, *SE* = 0.02, t(42) = -0.52, *p* = .61). These findings indicate no evidence that training produced a change in global response consistency.

weighting data. Specifically, standardized β values for spectral (β_Spec) and temporal (β_Dur) cues were calculated using z-scores for each of the 44 participants. A Euclidean distance matrix was computed, and Ward's minimum variance method was applied to derive clusters using the *hclust* function from the stats package in R (R Core Team, 2024). Visual inspection of the dendrogram (Figure 4) identified the largest vertical jump in linkage height at approximately six, indicating a natural division into three stable clusters. These clusters were visualized in a scatter plot (Figure 5a), with centroid analysis confirming distinct cue-weighting profiles across groups. Cluster 1 (*N* = 15) showed strong reliance on duration (Mean β_Dur = 2.37; Mean β_Spec = 0.09), consistent with a duration-dominant strategy. Cluster 2 (*N* = 22) showed moderate reliance on both the spectral and temporal cues but leaned more toward duration (Mean β_Dur = 0.90; Mean β_Spec = 0.30). Cluster 3 (*N* = 7) exhibited a spectral-dominant pattern (Mean β_Spec = 1.62; Mean β_Dur = 0.16).

[Insert Figure 4 around here]

[Insert Figure 5 around here]

To validate the cluster structure and explore the underlying dimensionality of cue-weighting behaviors, a principal component analysis was performed on the same β_Spec and β_Dur values. The principal component analysis revealed two principal components accounting for 100% of the variance: PC1 (82%) reflected the trade-off between spectral and temporal reliance (β_Spec = -0.71, β_Dur = 0.71), while PC2 (18%) represented the overall magnitude of cue use (β_Spec = 0.71, β_Dur = 0.71). When mapped onto the principal component analysis space (Figure 5b), the three clusters from hierarchical cluster analysis aligned along PC1 in a manner consistent with their centroid profiles: the duration-dominant group (Cluster 1) appeared on the right end of PC1, reflecting strong temporal weighting and low spectral weighting; the spectral-dominant group (Cluster 3) appeared on the left, indicating high spectral but low temporal reliance; and the duration-leaning balanced group (Cluster 2) clustered around the center of PC1, reflecting more balanced weighting. Along PC2, both the duration-dominant group (Cluster 1) and the spectral-dominant group (Cluster 3) showed higher values, indicating greater overall cue utilization. In contrast, the duration-leaning balanced group

(Cluster 2) occupied a lower position along this axis, suggesting a generally weaker cue reliance at both acoustic dimensions. Taken together, these findings confirm the presence of three distinct cue-weighting profiles at baseline. Based on their cue reliance patterns, the clusters were labeled as 'duration-dominant' (Cluster 1), 'duration-leaning balanced' (Cluster 2), and 'spectral-dominant' (Cluster 3).

To investigate whether participants' cue profiles at pretest (i.e., duration-dominant, duration-leaning balanced, and spectral-dominant) were predicted by the four individual background factors (Instruction, Experience, LProf, and OProf), Fisher's Exact Tests were conducted for categorical predictors (Instruction, Experience) and Kruskal–Wallis rank sum tests were conducted for continuous predictors (LProf and OProf). Consistent with the analysis reported in 3.1.1B, these analyses revealed no significant associations (full details of these results are provided in the Supplementary Materials).

3.2.2. Individual differences in cue-weighting change patterns

As in the pretest, individual regression coefficients for the spectral (β_Spec) and temporal (β_Dur) cues were extracted for each participant from separate mixed-effects models fitted to the control and experimental groups in the posttest, using the *coef* function in R. Based on these coefficients, a weighting score was computed for each participant using Formula (2) described in Section 2.4. To further examine the effect of the intervention on cue-weighting profiles, change scores were calculated for each participant by subtracting pretest values from posttest values for the regression coefficients of spectral reliance (Δβ_Spec), temporal reliance (Δβ_Dur), and the overall weighting score (Δweighting score). Figure 6 visualizes the distributions of these change scores across the two groups.

[Insert Figure 6 around here]

To examine whether participants exhibited distinct or universal patterns in their change of reliance on spectral and temporal cues, a hierarchical cluster analysis was conducted on Δβ_Spec and Δβ_Dur values from 44 participants. The resulting dendrogram (Figure 7) revealed two stable clusters

686 at a linkage height at 8. These clusters were projected onto a scatter plot (Figure 8), and centroid
687 analysis confirmed distinct patterns of cue change. Cluster 1 (*N* = 22), centered closely around zero
688 on both dimensions, showed minimal overall change—particularly negligible change in spectral
689 weighting (Mean Δβ_Dur = 0.14; Mean Δβ_Spec = –0.01). In contrast, Cluster 2 (*N* = 20) exhibited a
690 clear trade-off pattern, with a marked increase in spectral cue weighting and a concurrent decrease
691 in duration cue weighting (Mean Δβ_Spec = 1.08; Mean Δβ_Dur = –0.77).

692 [Insert Figure 7 around here]

693 [Insert Figure 8 around here]

694 Because the intervention did not shift cue-weighting patterns uniformly, we further examined
695 which factors predicted individual differences in change. We focused on whether the three change
696 scores (Δβ_Spec, Δβ_Dur and Δweighting score) were associated with 1) background variables
697 (Instruction, Experience, LProf, OProf) and 2) baseline cue use (pretest values of β_Dur, β_Spec and
698 weighting score). Only significant results are reported here; full model outputs are available in the
699 Supplementary Materials. First, to examine whether background variables predicted changes in cue
700 use, we conducted a series of linear regression analyses. We first tested the interaction between
701 Group (Experimental vs. Control) and each background variable (Experience, Instruction, OProf,
702 LProf) for Δβ_Dur, Δβ_Spec, and Δweighting score. For Δβ_Dur, there was a significant main effect of
703 LProf (β = –0.26, $t(37) = -2.16$, $p = .038$), with no interaction with Group ($p = .642$). This indicates that
704 higher self-rated listening proficiency predicted a greater decrease in reliance on duration cues, and
705 there was no evidence that the two groups differed. For Δβ_Spec, the analyses revealed a significant
706 Experience × Group interaction (β = 1.22, $t(37) = 2.50$, $p = .017$). Compared to the Control group,
707 Experimental participants with prior experience living in English-speaking environments showed
708 larger increases in spectral cue weighting after training. No other interactions were significant for
709 Δβ_Dur, Δβ_Spec, or Δweighting score. Given the significant Experience × Group interaction, we
710 conducted follow-up analyses within each group. In the Control group, Experience did not
711 significantly predict Δβ_Spec (β = –0.31, $t(16) = -0.69$, $p = .501$). In contrast, in the Experimental

group, Experience was a strong positive predictor of Δβ_Spec (β = 0.92, $t(21) = 3.61$, $p = .002$, adjusted $R^2 = .35$), indicating that for the Experimental group, participants with more prior experience in English-speaking contexts exhibited greater increases in spectral cue reliance following training.

We next examined whether initial cue reliance moderated changes in cue weighting following training. Linear regression models were fitted with the formula: Δβ_Dur / Δβ_Spec / Δweighting score ~ β_Spec_Pre × β_Dur_Pre × Group. No significant interactions were found for Δβ_Spec or Δweighting score. For Δβ_Dur, however, two significant effects emerged. First, there was a significant interaction between pretest duration reliance and Group (β_Dur_Pre × Group: β = −0.77, $p = .010$), indicating that participants in the experimental group showed a larger decrease in reliance on duration compared to those in the control group, consistent with the group-level results reported in Section 3.1.2B. Second, a significant interaction was found between pretest spectral and durational weighting (β_Spec_Pre × Dur_Pre: β = −0.71, $p = .017$), but this effect did not further interact with Group (β_Spec_Pre × Dur_Pre × Group: β = 0.10, $p = .862$). This pattern suggests that, across both groups, participants who initially relied heavily on both spectral and temporal cues were more likely to reduce their use of duration, with no evidence of group-specific differences in this tendency.

4 Discussion

This study first examined Mandarin listeners' baseline cue-weighting strategies of English /i:/-/ɪ/ at both group level and individual level. Pretest results showed that L1-Mandarin English learners relied on both the spectral and temporal cues at the group level, with vowel duration being the primary cue, although hierarchical cluster analysis and principal component analysis showed substantial individual variability in the individual cue-weighting patterns regarding both cue dominance and overall magnitude of cue use. Critically, we also investigated whether Mandarin-L1 English learners' cue-weighting patterns of English /i:/-/ɪ/ could be shifted via explicit phonetic instruction that directly raised metalinguistic awareness of the function of the two acoustic dimensions. In contrast to Zheng (2023), no correlational evidence was found between cue use

patterns at pretest and participants' self-reported prior phonetic instruction. However, critically, group level comparison of the experimental and control groups indicated that explicit instruction reduced reliance on vowel duration at posttest, while both groups' reliance on the spectral cue increased. However, there was systematic individual variability in how training benefited listeners. In the following discussion, we start by considering the pretest findings in the light of the previous literature and then turn to consider the results of the intervention.

4.1 Baseline L2 cue-weighting patterns at pretest

Pretest data showed that duration was the primary cue used by Mandarin-L1 English listeners at the group level, which is consistent with previous findings (e.g., Bohn, 1995; Wang & Munro, 1999; Yang et al., 2016). However, the extent to which these learners rely on spectral cues varied across studies. While Bohn (1995) and Wang and Munro (1999) suggest an exclusive reliance on duration, the present study and Yang et al. (2016) indicate participants do rely on vowel spectrum, though only as a secondary cue. Such inconsistency might stem from differences in sampling. Participants in Yang et al. (2016) and the current research were university students, who likely have higher English proficiency and greater exposure to authentic English input than participants in earlier studies such as Bohn (1995), who had very limited exposure to native English speakers. According to the L2LP model, learners with sufficient input may begin to recognize the reliability of spectral cues and form new categories along this dimension, thus progressing from Stage 1 (duration-only) to Stages 2 and 3, where spectral cues become increasingly important. Thus, the variation in findings may reflect differences in learners' learning stages, with Yang et al. (2016) and the present study including more participants at advanced stages.

Consistent with this interpretation, our individual-level analyses reveal systematic variation across learners that maps onto these different stages of perceptual learning. Hierarchical cluster analysis of participants' β_Dur and β_Spec values revealed three distinct subgroups corresponding to these three different stages of perceptual learning proposed in the L2LP model (Escudero, 2000, 2005). The first group, duration-dominant (Mean β_Dur = 2.37; Mean β_Spec = 0.09), showed strong

reliance on duration with minimal spectral weighting, corresponding to Stage 1, where learners rely solely on duration cues. The second group, duration-leaning balanced (Mean β_Dur = 0.90; Mean β_Spec = 0.30), on average showed a more balanced cue use pattern (though still duration-leaning). This is consistent with Stage 2, where spectral sensitivity begins to emerge while duration still dominates. The third group, spectral-dominant (Mean β_Spec = 1.62; Mean β_Dur = 0.16), relied primarily on spectral cues, using duration only secondarily, as predicted for Stage 3 learners. To further assess the magnitude of cue use across groups, we conducted a principal component analysis. The second principal component (PC2), which captured the overall strength of cue reliance, revealed an interesting pattern: learners in the duration-dominant (Stage 1) and spectral-dominant (Stage 3) groups showed stronger cue use overall, whereas the duration-leaning balanced group (Stage 2) exhibited generally weaker weighting on both dimensions. This intermediate-level cue reliance magnitude may reflect a transitional 'trade-off' period in perceptual development, where spectral weighting is still emerging while duration is being down-weighted. Similar patterns were observed in Morrison (2008) for Spanish-speaking learners of English, which suggests that this intermediate dip in overall cue strength may capture the dynamic nature of the progressive perceptual learning stages.

One further implication of our findings is that they challenge Morrison's (2008) refinement of the L2LP model. Morrison observed a subset of Spanish learners who relied on both spectral and temporal cues, but in a counterintuitive pattern: they tended to label longer vowels that were spectrally /ɪ/-like as /i:/, and shorter vowels that were spectrally /i:/-like as /ɪ/. To explain this, he proposed a transitional 'Stage 1/2,' arguing that learners' use of duration reflects a multidimensional 'category-goodness-difference assimilation' process. Specifically, (1) English orthographic 'i' often corresponds to /ɪ/, while Spanish 'i' maps to Spanish /i/, which may link English /ɪ/ to Spanish /i/; and (2) Spanish /i/ is spectrally close to English /i:/ but durationally similar to /ɪ/. As a result, learners may categorize shorter, Spanish-/i/-like tokens as English /ɪ/ and longer, spectrally dissimilar tokens as /i:/, making duration a more reliable cue than spectrum during the early stages of learning. This account

790 diverges from L2LP's explanation in a critical way. Whereas L2LP attributes reliance on duration to its
791 status as a 'blank slate' cue (unused contrastively in the L1 and therefore easier to recruit), Morrison
792 suggests that it is because both spectral and temporal properties are already encoded in their L1
793 category. Applied to Mandarin, however, this logic would predict initial reliance on vowel spectrum:
794 Mandarin /i/ closely matches English /i:/ spectrally, while its duration is highly context-dependent,
795 which makes it neither inherently long nor short and provides no consistent anchor. L1-Mandarin
796 English learners should therefore map English /i:/ to Mandarin /i/ and categorize spectrally poor
797 matches as /ɪ/, beginning at Stage 3 with spectral reliance. Yet our data indicate otherwise—many
798 Mandarin learners relied almost exclusively on duration at pretest. Therefore, our findings align more
799 closely with L2LP's 'blank slate' hypothesis, suggesting that Morrison's account may reflect patterns
800 specific to Spanish learners, whereas L2LP provides a broader explanation of cue weighting across L1
801 backgrounds.
802 This study also examined whether the correlation between prior phonetic instruction and
803 spectral cue use reported in Experiment 1 of Zheng (2023) would replicate. Unlike Zheng's (2023)
804 findings, there was no evidence of such a correlation , despite using the same questionnaire and
805 task. This inconsistency may reflect random variation between samples: it remains unclear whether
806 the earlier finding represents a true effect or a Type I error, or whether the current result a Type II
807 error. More importantly, both studies relied on participants' retrospective reports, which may have
808 been influenced by their task performance and are therefore not a reliable basis for drawing
809 conclusions. These limitations suggest the need to move beyond correlational approaches and adopt
810 a more experimental method via intervention. The present study addressed this by directly giving
811 explicit phonetic instructions through a controlled intervention, which could allow us to evaluate the
812 causal impact of explicit phonetic instruction on L2 cue weighting.
813 **4.2 The effects of explicit phonetic instruction**
814 ***4.2.1 Effects at the group level***

The present study provides causal evidence that explicit instruction can influence the perceptual cue-weighting behaviors of L1-Mandarin English learners: at the group level, the key finding is that participants who were explicitly informed that vowel duration is not a reliable cue on its own significantly reduced their reliance on this dimension, whereas the control group showed no such change. This suggests that explicit phonetic training can effectively help Mandarin-speaking learners reduce their dependence on duration when perceiving the English /iː/–/ɪ/ contrast. This finding has implications for the L2LP model, which specifies distributional learning through exposure to input as the mechanism underlying cue-weighting development. The current results demonstrate that explicit instruction alone can also facilitate changes in cue weighting—indicating that distributional learning may not be the sole path to perceptual adjustment. However, in line with the L2LP hypothesis, the effectiveness of explicit instruction in reducing duration reliance may rely on the 'blank slate' nature of vowel duration for Mandarin learners: Because vowel duration is not used contrastively in Mandarin, learners may be more cognitively flexible in adjusting their reliance on this cue when guided by explicit instruction. Similar conclusions were drawn by Kondaurova and Francis (2010), who successfully reduced Spanish-speaking learners' reliance on duration using inhibitory training that introduced high variability along the duration dimension, making it less informative for categorization. These findings suggest that for 'blank slate' dimensions, explicit instruction alone may be sufficient to induce cue reweighting, without the need for extensive input. However, it is possible that explicit instruction may be less effective when the target dimension is instead entrenched in the L1. For instance, Jasmin et al. (2021) found that while L1-Spanish speakers were able to ignore pitch (the dominant cue) and to focus on amplitude when explicitly instructed to do so, L1-Mandarin speakers struggled to do so, likely because Mandarin is a tonal-language where pitch is a primary

carrier of lexical meanings. This emphasizes the influence of L1 experience on cue weighting and suggests that explicit instruction may be insufficient when it conflicts with dominant L1 cue strategies.

Furthermore, an unexpected result was that both groups demonstrated an increase in spectral cue weighting after the first session, with no evidence that the degree of increase differed between the two groups. The increase in spectral cue weighting in the control group was possibly due to the intensive exposure to spectral variability during the pre- and posttests. Since the stimuli varied extensively along the spectral dimension, listeners may have recognized this variability as a 'potential source of information' (Holt & Lotto, 2006, p. 3068) which led to increased attention to vowel quality and increased spectral reliance. This interpretation echoes previous findings that passive exposure alone can shift cue weighting by drawing attention to the less-weighted cue (e.g., Escudero et al., 2011; Holt & Lotto, 2006; Kachlicka et al., 2024; Wanrooij et al., 2013). One possible reason why the experimental and control groups did not differ significantly in their increase in spectral cue weighting could be that the increase in spectral cue weighting in both groups is due to exposure to spectral variability rather than the explicit instruction. Alternatively, it could be that explicit instruction did contribute to the shift in spectral cue weighting, but ceiling effects limit how much change can occur. Future studies should aim to disentangle the respective contributions of passive exposure and explicit instruction in shifting cue-weighting behaviors.

4.2.2 Individual differences in the benefits of explicit instruction

Although explicit instruction in cue weighting led to overall changes in the experimental group, not all participants benefited equally, which suggests that certain learner-specific factors might have influenced the degree to which instruction facilitates cue-weighting adjustment. Firstly, for both control and experimental groups, listeners with higher self-rated listening proficiency tended to show a greater decrease in reliance on duration cues after intervention. This might indicate more proficient listeners were better able to shift away from a less reliable cue and that lower-proficiency listeners

relied more rigidly on duration as a strategy and were less able to change their perceptual weighting, even after training.

The second factor influencing cue reallocation was participants' initial cue use prior to training. Specifically, in comparison with the control group, individuals in the experimental group who relied more heavily on duration at pretest tended to show larger reductions in duration reliance after treatment. This is not unexpected because those with a stronger initial dependence on duration had more room for change. In addition, across both groups, participants who initially weighted both duration and spectral cues more heavily exhibited even greater reductions in duration reliance. This effect did not interact with group, indicating that learners with a broader initial sensitivity to acoustic cues were generally more flexible in reallocating attention across dimensions.

The third–and probably most important—factor that influences the benefits of training was the amount of English-language experience. This factor was the only one that showed a significant *Group* × *Test* interaction, indicating that its effect was specific to the experimental group. In the experimental group, participants who had experience living in English-speaking environments showed greater increases in spectral cue reliance after training. This suggests existing exposure to rich L2 input may facilitate learners' ability to process and internalize explicit information about phonetic cues. In other words, instruction appears to be more effective when learners have already developed a foundational understanding of the L2 sound system through input. This observation has broader theoretical implications. While models such as L2LP emphasize input-driven distributional learning as the foundation of cue-weighting development, the present findings indicate that instruction can play a complementary role. However, its effectiveness may still depend on prior exposure to L2 input. Explicit phonetic instruction, in this sense, may function not as a substitute for input, but as a mechanism that helps learners make better use of the input they have received. Therefore, theoretical models of L2 speech learning like L2LP may need to account for how instructional interventions interact with learners' input histories to shape perceptual development.

4.2.3 Pedagogical implications

The present findings have several implications for second language phonetic learning. First, they provide evidence that explicit phonetic instruction can successfully shift learners' perceptual cue weighting, at least for 'blank slate' dimensions such as vowel duration in Mandarin learners of English. In such cases, explicit instruction offers a direct and efficient means of guiding attention to the functional role of the cue, comparable in effectiveness to input-based or error-driven approaches (e.g., Kondaurova & Francis, 2010; Zhang et al., 2021). However, it is worth mentioning whether the effects of explicit instruction generalize to other L1 backgrounds, other cue dimensions, or yield durable learning outcomes remains an open question. For instance, for 'non-blank slate' dimensions, such as vowel spectrum in this study, explicit instruction alone may be insufficient. A practical recommendation is therefore to combine explicit instruction with input-rich (e.g., bimodal distributional learning) or error-driven (e.g., HVPT) training paradigms to achieve more robust perceptual change.

Second, the effectiveness of explicit instruction appears to depend on learners' prior L2 experience. In this study, instruction was most beneficial for learners who had lived in English-speaking environments, suggesting that exposure to rich input helps learners process and internalize the explicit instruction. Pedagogically, this implies that explicit instruction may be particularly useful for 'blank slate' cues when learners already have rich L2 exposure. More generally, combining explicit instruction with rich L2 input may maximize learning outcomes.

Finally, it is important to acknowledge that not all acoustic cues are easily verbalizable or teachable through explicit instruction. Some contrasts such as English /r/–/l/ for Japanese learners, which involve subtle spectral properties like F2/F3 transitions, may be less accessible to conscious awareness. Previous research has shown that such contrasts can sometimes be learned more efficiently through incidental, input-driven exposure rather than explicit instruction (e.g., Lim & Holt, 2011). Second language teachers should therefore consider both the nature of the cue and the learners' L1 backgrounds when designing phonetic training.

4.3 Limitations and future directions

We would also like to acknowledge several limitations of the present study. First, our sample size is relatively small, with 20 participants in the control group and 24 in the experimental group. As a result, findings (particularly those concerning individual differences) should be interpreted with caution and validated in future work with larger samples. Second, the study focused on cue-weighting patterns rather than categorization performance, such as identification accuracy. Future research should further investigate whether changes in cue-weighting strategies can lead to improved perception of the contrast. Third, the present study relied on self-rated proficiency measures to characterize participants' English proficiency. Future studies should complement such self-reported data with standardized proficiency assessments, such as the LexTALE (Lemhöfer & Broersma, 2012), to provide a more objective evaluation of participants' language proficiency levels. Finally, participants showed a general bias towards categorizing stimuli as 'bead', with the experimental group exhibiting a stronger bias than the control group at pretest. However, the absence of Test and Group × Test interactions indicates no evidence that training altered this bias within or between groups. The baseline group difference likely reflects accidental sampling variation. Several factors may also account for the overall 'bead' bias. Because Mandarin has only one high front vowel /i/, which is acoustically closer to English /i:/ than /ɪ/, Mandarin listeners may have been more inclined to select 'bead.' Properties of the stimulus continuum may have also played a role. For instance, if the duration range aligned more closely with participants' internal representation of /i:/, this could have nudged responses toward 'bead' without reflecting a genuine perceptual preference. Importantly, however, this baseline difference and general bias do not undermine the main findings in the current research, as all tests of intervention effectiveness were based on changes from pretest to posttest rather than absolute response proportions.

5 Conclusion

This study investigated the role of explicit phonetic instruction in reshaping Mandarin-speaking learners' cue weighting for the English /i:/–/ɪ/ contrast and revealed three key findings. First, learners relied predominantly on duration at the group level prior to intervention, with individual variability

patterns consistent with the different stages of perceptual learning proposed by the L2LP model, supporting its cross-linguistic generalizability. Second, explicit instruction was effective in reducing reliance on duration, showing that explicit instruction with a metalinguistic nature can complement input-driven learning by directing shifting learners' attention to the functional role of acoustic cues. Finally, training outcomes were influenced by individual differences: learners with stronger listening proficiency, higher baseline cue sensitivity, and prior experience in English-speaking environments benefited most. Together, these findings indicate explicit instruction is a valuable tool for promoting perceptual learning, particularly for the 'blank slate' acoustic dimensions which are not used contrastively in learners' L1, while highlighting the importance of prior L2 exposure in shaping instructional effectiveness.

6 Acknowledgments

This research was supported by the Department of Education at University of Oxford, where the original study was conducted as part of the first author's master's thesis, and by research funds provided by St John's college to the last author.

7 Data Availability Statement

The datasets generated during and/or analyzed during the current study are available in the OSF repository, https://osf.io/r9qbk/?view_only=571141f6cf514c3082bfbabbe09fd4ba

References

Andringa, S. (2020). The emergence of awareness in uninstructed L2 learning: A visual world eye tracking study. *Second Language Research*, *36*(3), 335–357.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Baek, H. (2022). Prosodic cue weighting in the processing of ambiguous sentences by native English

965 listeners and Korean listeners of English. *The Journal of the Acoustical Society of America*, *151*(1),
966 158–167. https://doi.org/10.1121/10.0009171
967 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**.
968 *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01
969 Best, C. T. (1995). A direct realist view of cross-language speech perception. *Speech Perception and*
970 *Linguistic Experience*, *171*.
971 Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* (Version Version 6.3.14)
972 [Computer software]. http://www.praat.org/
973 Bohn, O.-S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In
974 *Speech perception and linguistic experience: Issues in cross-language research* (pp. 279–304).
975 York Press.
976 Chandrasekaran, B., Yi, H.-G., Smayda, K. E., & Maddox, W. T. (2016). Effect of explicit dimensional
977 instruction on speech category learning. *Attention, Perception, & Psychophysics*, *78*(2), 566–582.
978 https://doi.org/10.3758/s13414-015-0999-x
979 Chen, H. C., & Tian, J. (2020). The Effects of Explicit Rule and Acoustic-perceptual Instructions on Chinese
980 ESL Learners' Prosodic Acquisition of English Lexical Stress. *Speech Prosody 2020*, 833–837.
981 https://doi.org/10.21437/SpeechProsody.2020-170
982 Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-
983 Chinese. *The Journal of the Acoustical Society of America*, *123*(4), 2356–2366.
984 https://doi.org/10.1121/1.2839004
985 Clayards, M. (2018). Differences in cue weights for speech perception are correlated for individuals within
986 and across contrasts. *The Journal of the Acoustical Society of America*, *144*(3), EL172–EL177.
987 https://doi.org/10.1121/1.5052025

Colantoni, L., Steele, J., & Escudero, P. R. (2015). *Second language speech: Theory and practice*. Cambridge University Press.

Conover, L. (2023). The direction of attention in second language phonological contrast learning. *The Journal of the Acoustical Society of America*, *153*(6), 3390. https://doi.org/10.1121/10.0019714

Crystal, T. H., & House, A. S. (1988). The duration of American-English vowels: An overview. *Journal of Phonetics*, *16*(3), 263–284. https://doi.org/10.1016/S0095-4470(19)30500-5

Escudero, P. (2000). *Developmental patterns in the adult L2 acquisition of new contrasts: The acoustic cue weighting in the perception of Scottish tense/lax vowels by Spanish speakers* [Unpublished Master's thesis]. University of Edinburgh.

Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization* [PhD Thesis]. Utrecht University & LOT.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics*, *37*(4), 452–465. https://doi.org/10.1016/j.wocn.2009.07.006

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America*, *130*(4), EL206–EL212. https://doi.org/10.1121/1.3629144

Escudero, P., & Boersma, P. (2004). BRIDGING THE GAP BETWEEN L2 SPEECH PERCEPTION RESEARCH AND PHONOLOGICAL THEORY. *Studies in Second Language Acquisition*, *26*(4), 551–585. https://doi.org/10.1017/S0272263104040021

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, *92*(1), 233–277.

Flege, J. E., Bohn, O.-S., & Jang, S. (1997). Effects of experience on non-native speakers' production and

perception of English vowels. *Journal of Phonetics*, *25*(4), 437–470.

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234–1251. https://doi.org/10.1121/1.2945161

Francis, A. L., & Nusbaum, H. C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349–366. https://doi.org/10.1037/0096-1523.28.2.349

Giannakopoulou, A., Brown, H., Clayards, M., & Wonnacott, E. (2017). High or low? Comparing high and low-variability phonetic training in adult and child second language learners. *PeerJ*, *5*, e3209. https://doi.org/10.7717/peerj.3209

Grenon, I., Kubota, M., & Sheppard, C. (2019). The creation of a new vowel category by adult learners after adaptive phonetic training. *Journal of Phonetics*, *72*, 17–34. https://doi.org/10.1016/j.wocn.2018.10.005

Guion, S. G., & Pederson, E. (2007). Investigating the role of attention in phonetic learning. In O.-S. Bohn & M. J. Munro (Eds), *Language Learning & Language Teaching* (Vol. 17, pp. 57–77). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.17.09gui

Hao, Y.-C. (2018). Second Language Perception of Mandarin Vowels and Tones. *Language and Speech*, *61*(1), 135–152. https://doi.org/10.1177/0023830917717759

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071.

https://doi.org/10.1121/1.2188377

Holt, L. L., Tierney, A. T., Guerra, G., Laffere, A., & Dick, F. (2018). Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing. *Hearing Research*, *366*, 50–64. https://doi.org/10.1016/j.heares.2018.06.014

Icardo Isasa, A. (2020). *The perception, production and the effect of explicit instruction in L2 English and Spanish vowels in foreign language (FL) adult learners* [PhD Thesis]. University of Illinois at Urbana-Champaign.

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, *132*(6), 3950–3964. https://doi.org/10.1121/1.4765076

Ingvalson, E. M., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics*, *39*(4), 571–584. https://doi.org/10.1016/j.wocn.2011.03.003

Jasmin, K., Sun, H., & Tierney, A. T. (2021). Effects of language experience on domain-general perceptual strategies. *Cognition*, *206*, 104481. https://doi.org/10.1016/j.cognition.2020.104481

Kachlicka, M., Symons, A. E., Ruan, Y., Saito, K., Dick, F., & Tierney, A. (2024). *Effects of targeted perceptual training on L2 suprasegmental weighting strategies*.

Kajouj, F., & Kager, R. (2019). Effects of bilingualism on cue weighting: How do bilingual children perceive the Dutch [ɑ]-[a:] contrast? *International Journal of Bilingualism*, *23*(2), 509–524. https://doi.org/10.1177/1367006917745696

Kivistö-de Souza, H., Carlet, A., Julkowska, I., & Rato, A. (2017). Vowel inventory size matters: Assessing cue-weighting in L2 vowel perception. *Ilha Do Desterro A Journal of English Language, Literatures in English and Cultural Studies*, *70*(3), 33–46. https://doi.org/10.5007/2175-8026.2017v70n3p33

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857. https://doi.org/10.1121/1.398894

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics*, *38*(4), 569–587. https://doi.org/10.1016/j.wocn.2010.08.003

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57. https://doi.org/10.1016/j.wocn.2016.08.006

Kong, E. J., & Kang, S. (2023). Individual Differences in Categorical Judgment of L2 Stops: A Link to Proficiency and Acoustic Cue-Weighting. *Language and Speech*, *66*(2), 354–380. https://doi.org/10.1177/00238309221108647

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lim, S., & Holt, L. L. (2011). Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. *Cognitive Science*, *35*(7), 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

Lipski, S. C., Escudero, P., & Benders, T. (2012). Language experience modulates weighting of acoustic cues for vowel perception: An event-related potential study. *Psychophysiology*, *49*(5), 638–650. https://doi.org/10.1111/j.1469-8986.2011.01347.x

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a Tonal Language by Attending to the Tone: An In Vivo Experiment. *Language Learning*, *61*(4), 1119–1141.

https://doi.org/10.1111/j.1467-9922.2011.00673.x

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. https://doi.org/10.1016/S0010-0277(01)00157-3

Morrison, G. S. (2008). L1-Spanish Speakers' Acquisition of the English /i /—/I/ Contrast: Duration-based Perception is Not the Initial Developmental Stage. *Language and Speech*, *51*(4), 285–315. https://doi.org/10.1177/0023830908099067

Nishi, K., & Kewley-Port, D. (2007). Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1496–1509. https://doi.org/10.1044/1092-4388(2007/103)

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *The Journal of the Acoustical Society of America*, *115*(4), 1777–1790. https://doi.org/10.1121/1.1651192

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ruan, Y., & Saito, K. (2023). Less precise auditory processing limits instructed L2 speech learning: Communicative focus on phonetic form revisited. *System*, *114*, 103020. https://doi.org/10.1016/j.system.2023.103020

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, & Psychophysics*, *78*(1), 355–367. https://doi.org/

10.3758/s13414-015-0987-1

Tremblay, A., Kim, H., Kim, S., & Cho, T. (2023). Perceptual training enhances the use of vowel quality cues to lexical stress: The benefits of intonational variability. *Proceedings of the 20th International Congress of Phonetic Sciences*.

Udovchenko, I. (2016). *L1 influence on acoustic cue weighting in perception and production of English high front vowel contrasts* [Doctoral]. California State University, Long Beach.

Wang, X., & Munro, M. J. (1999). The perception of English tense-lax vowel pairs by native Mandarin speakers: The effect of training on attention to temporal and spectral cues. *Proceedings of the 14th International Congress of Phonetic Sciences*, *3*, 125–128.

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, *32*(4), 539–552. https://doi.org/10.1016/j.system.2004.09.011

Wanrooij, K., Escudero, P., & Raijmakers, M. E. J. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. *Journal of Phonetics*, *41*(5), 307–319. https://doi.org/10.1016/j.wocn.2013.03.005

Wiener, S., Murphy, T. K., & Holt, L. L. (2025). Incidental Nonspeech Auditory Learning Scaffolds Phonetic, Category, and Word Learning in a Foreign Language Classroom. *Language Learning*, lang.12700. https://doi.org/10.1111/lang.12700

Yang, X., Shi, F., Liu, X., & Zhao, Y. (2016). Learning styles and perceptual patterns for English /i/ and /ɪ/ among Chinese college students. *Applied Psycholinguistics*, *37*(3), 673–701. https://doi.org/10.1017/S014271641500020X

Yazawa, K., Whang, J., Kondo, M., & Escudero, P. (2020). Language-dependent cue weighting: An investigation of perception modes in L2 learning. *Second Language Research*, *36*(4), 557–581. https://doi.org/10.1177/0267658319832645

Yu, L. (2023). *The perceptual cue weighting of English tense-lax vowel contrasts by first language (L1) and second language (L2) speakers of English* [Doctoral]. Indiana University.

Zhang, X., Cheng, B., Qin, D., & Zhang, Y. (2021). Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics*, *87*, 101071. https://doi.org/10.1016/j.wocn.2021.101071

Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition*, *187*, 178–187. https://doi.org/10.1016/j.cognition.2019.03.004

Zheng, Z. (2023). *Acoustic cue-weighting in L1 and L2 English vowel perception* [University of Oxford]. https://doi.org/10.5287/ORA-BE4DYYN7R

Figure Captions

Figure 1. Distribution of vowels in the stimuli used for the word recognition task.

Figure 2. Visualization of response patterns for the experimental group and control group across the pre-test and post-test.

Figure 3. Distribution of β_Spec, β_Dur and weighting scores for the control group and experimental group in pretest.

Figure 4. Dendrogram of the hierarchical cluster analysis conducted on listeners' β_Spec and β_Dur in the pretest.

Figure 5. (a) Scatter plot showing the three clusters of cue weighting profiles identified through hierarchical cluster analysis based on pretest β_Spec and β_Dur; (b) principal component analysis scatter plot displaying the same three clusters in reduced dimensional space.

Figure 6. Distribution of Δβ_Spec, Δβ_Dur and Δweighting scores for the control group and experimental group.

Figure 7. Dendrogram of the hierarchical cluster analysis conducted on listeners' Δβ_Spec and Δβ_Dur.

Figure 8. Scatter plot showing the two clusters of change profiles identified through hierarchical cluster analysis based on Δβ_Spec and Δβ_Dur.

Table Titles

Table 1. Summary statistics (means, standard deviations, minimum values and maximum values) for self-rated English proficiency across four language skills and overall proficiency (N = 48).

Table 2. Output of mixed-effects logistic regression model examining the influence of Instruction, Experience, OProf and LProf on the cue weighting patterns of the 41 L2 English speakers in the pre-test.

Table 3. Statistical output of the mixed-effects logistic regression model to account for the test-group interaction.

Table 4. Statistical output of the mixed-effects logistic regression model for the experimental and control groups.

Figures

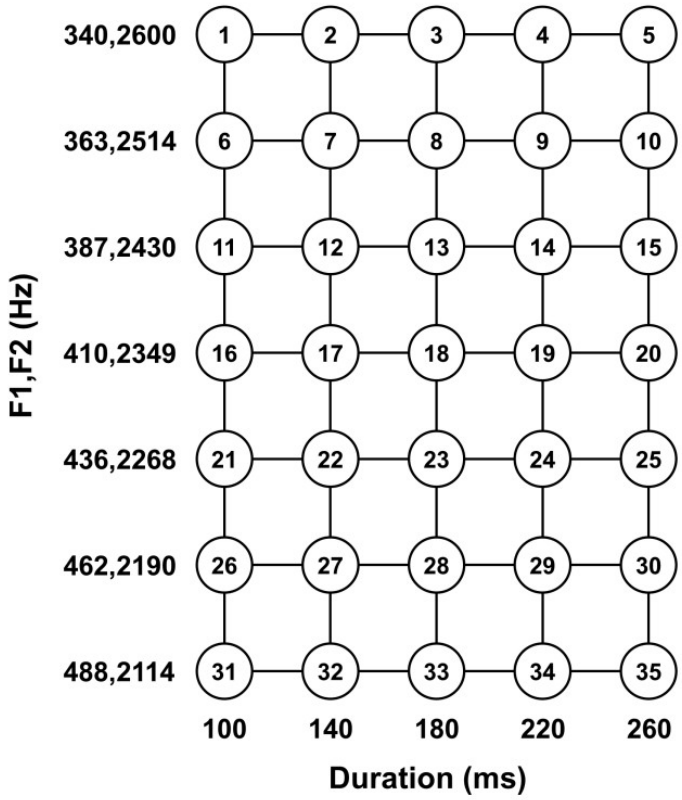


Figure 1.

Distribution of vowels in the stimuli used for the word recognition task.

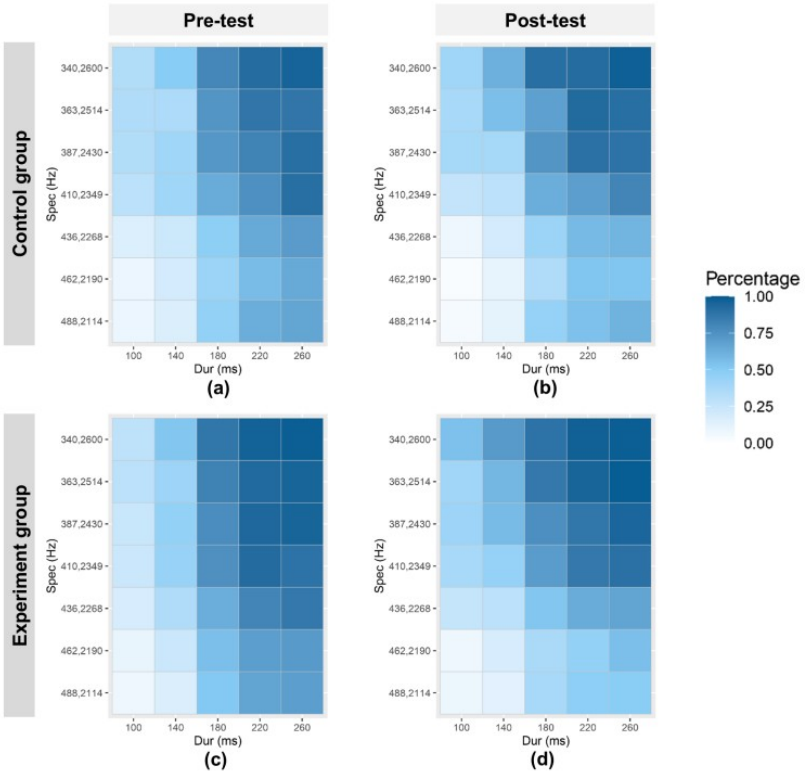


Figure 2.

Visualization of response patterns for the experimental group and control group across the pre-test and post-test

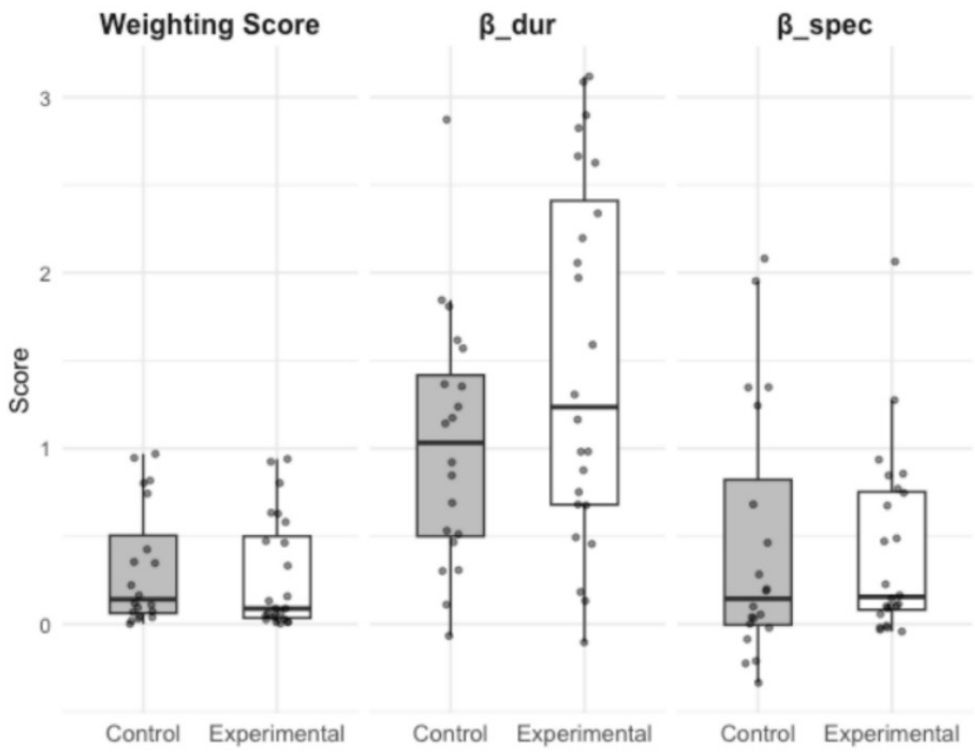


Figure 3.

Distribution of β_Spec, β_Dur and weighting scores for the control group and experimental group in pretest

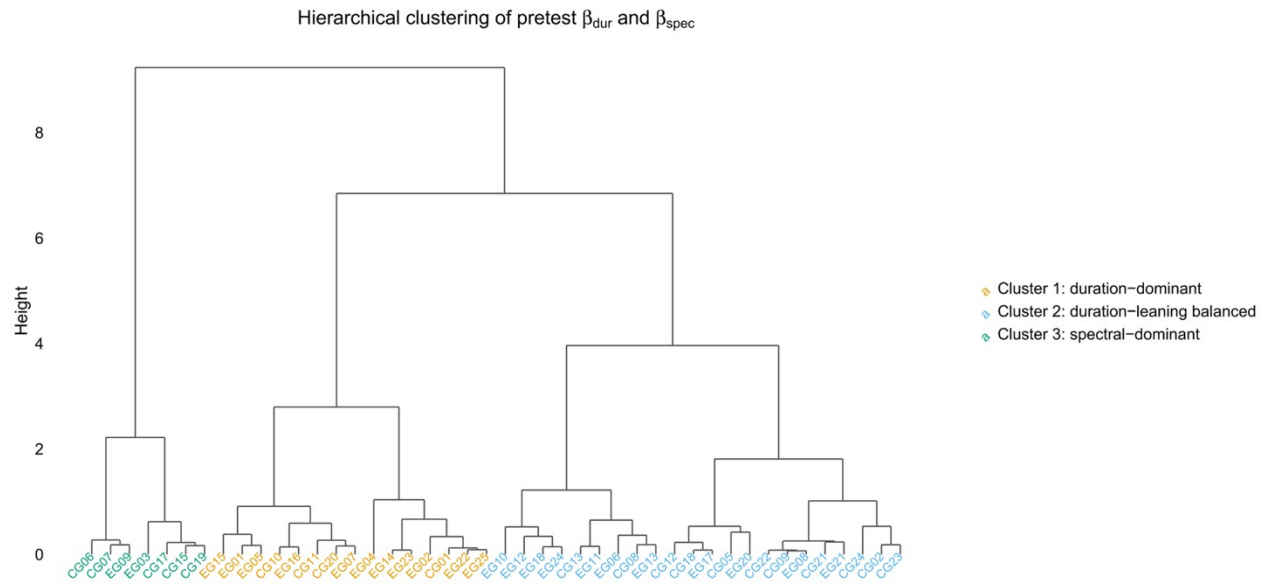


Figure 4.

Dendrogram of the hierarchical cluster analysis conducted on listeners' β_Spec and β_Dur in the pretest

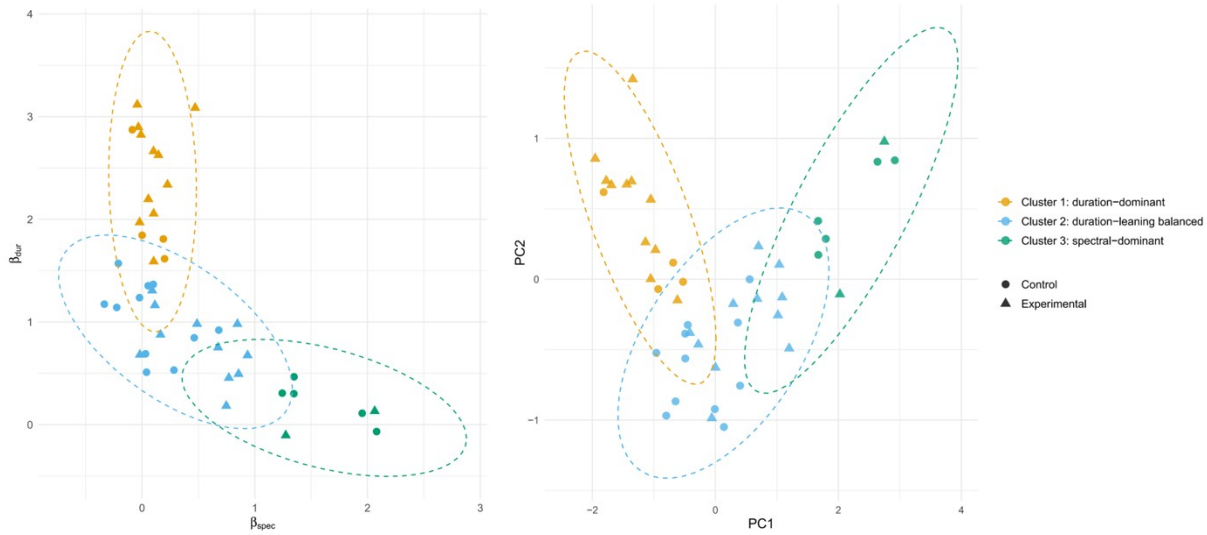


Figure 5.

(a) Scatter plot showing the three clusters of cue weighting profiles identified through hierarchical cluster analysis based on pretest β_Spec and β_Dur; (b) PCA scatter plot displaying the same three clusters in reduced dimensional space.

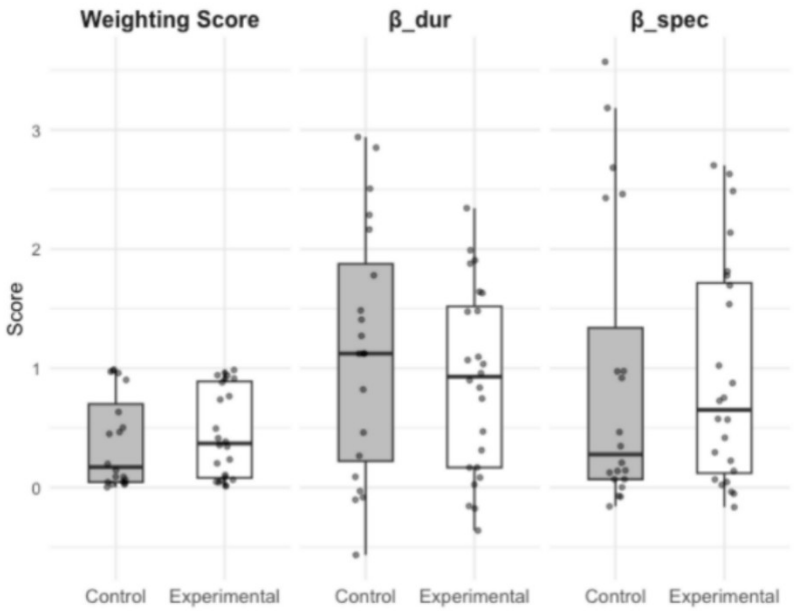


Figure 6.

Distribution of β_Spec, β_Dur and weighting scores for the control group and experimental group in post-test.

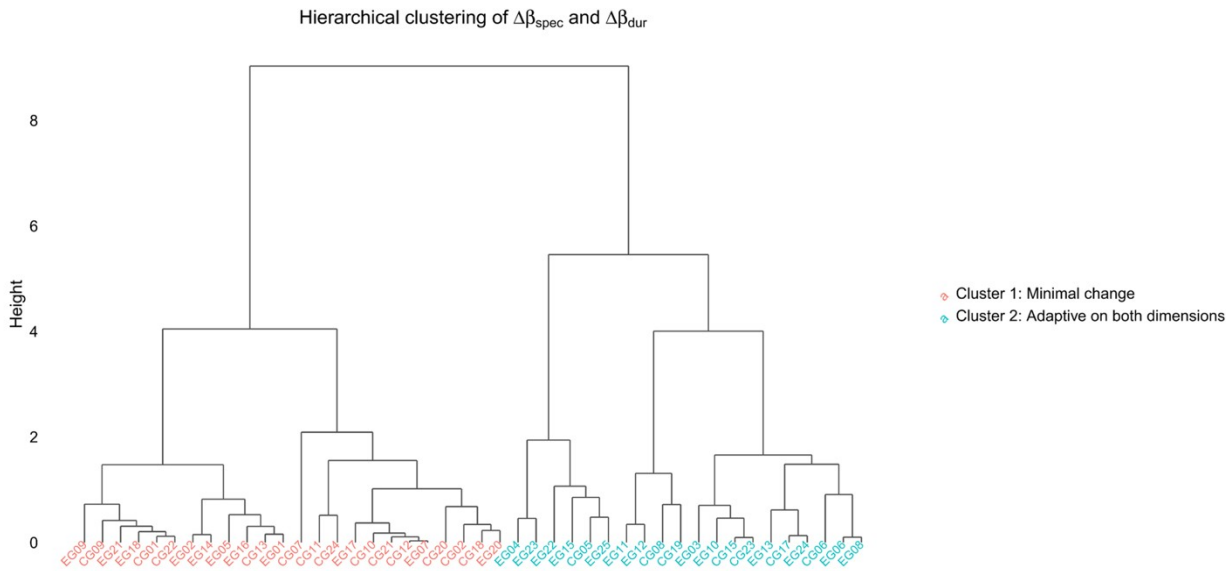


Figure 7.

Dendrogram of the hierarchical cluster analysis conducted on listeners' β_Spec and β_Dur in the post-test

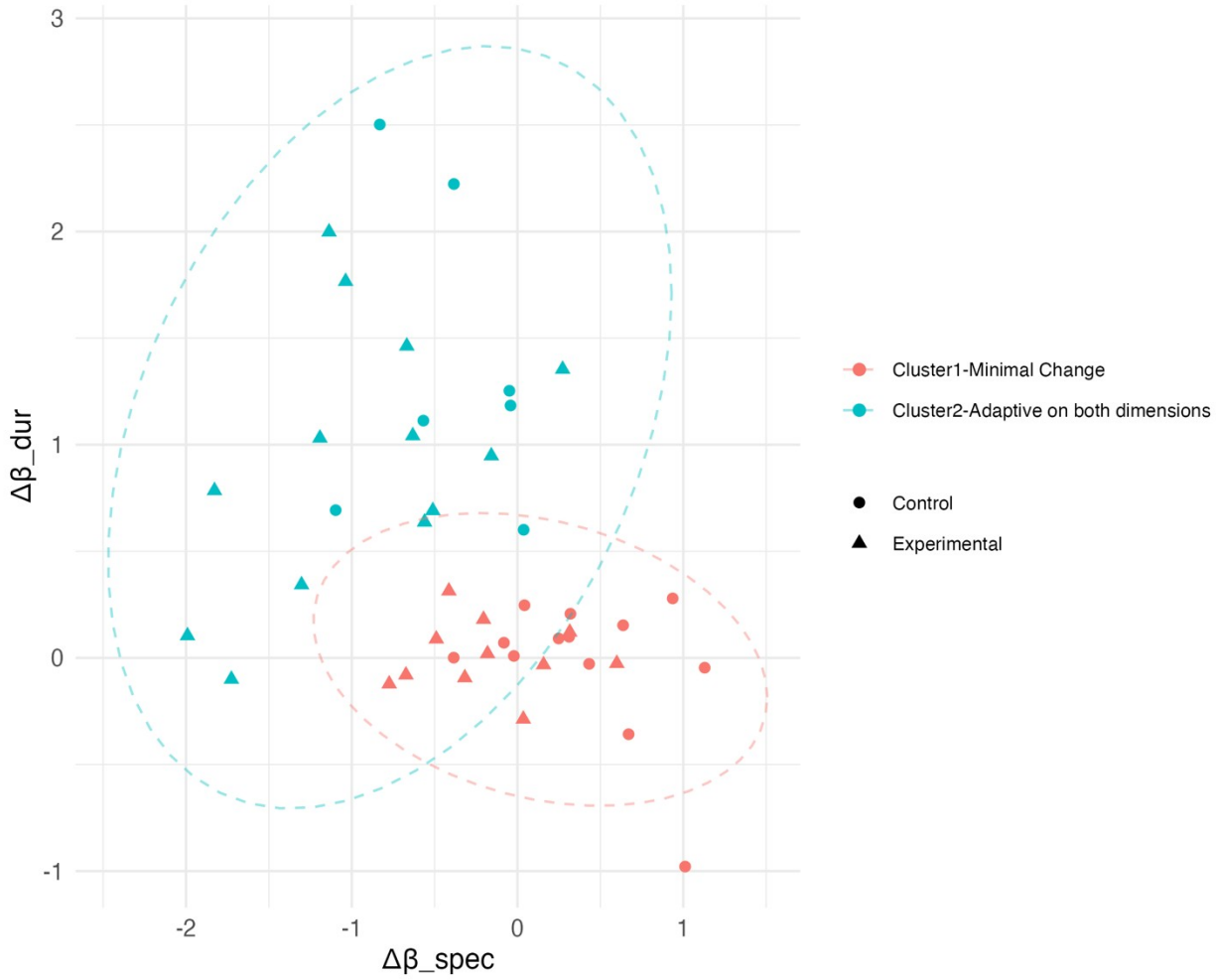


Figure 8. Scatter plot showing the two clusters of change profiles identified through hierarchical cluster analysis based on Δβ_Spec and Δβ_Dur.

Table 1.

Summary statistics (means, standard deviations, minimum values and maximum values) for self-rated English proficiency across four language skills and overall proficiency (N = 48).

Proficiency Measure	Mean	Standard Deviation	Minimum	Maximum
Speaking Proficiency	4.54	1.13	2	7
Listening Proficiency	4.90	0.97	3	7
Reading Proficiency	5.52	0.95	4	7
Writing Proficiency	4.73	0.98	3	7
Overall Proficiency	19.69	3.45	13	27

Tables

Table 2.

Output of mixed-effects logistic regression model examining the influence of Instruction, Experience, OProf and LProf on the cue weighting patterns of the 41 L2 English speakers in the pre-test.

Fixed Effect	**Estimate (β)**	**Standard Error**	*z*	*p*
(Intercept)	0.70	0.23	3.08	.002**
Spec	0.66	0.17	3.91	< .001***
Experience	−0.42	0.30	−1.43	.152
Dur	0.82	0.26	3.11	.002**
Instruction	0.09	0.26	0.37	.715
OProf	0.02	0.08	0.21	.834
LProf	−0.14	0.27	−0.54	.591
Spec × Experience	0.20	0.22	0.92	.358
Experience × Dur	−0.61	0.35	−1.77	.077†
Spec × Instruction	0.11	0.19	0.55	.580
Dur × Instruction	0.08	0.30	0.27	.786
Spec × OProf	−0.03	0.06	−0.47	.638
Dur × OProf	0.03	0.09	0.29	.773
Spec × LProf	−0.10	0.20	−0.50	.619
Dur × LProf	0.32	0.31	1.04	.297

Notes:

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

Table 3.

Statistical output of the mixed-effects logistic regression model to account for the test-group interaction.

Fixed Effect	**Estimate (β)**	**Standard Error**	***z***	***p***
(Intercept)	0.66	0.11	6.02	< .001***
Spec	0.69	0.12	5.56	< .001***
Group	0.49	0.22	2.25	.025*
Test	−0.01	0.14	−0.06	.955
Dur	1.17	0.14	8.56	< .001***
Spec × Group	0.00	0.25	0.02	.988
Spec × Test	0.51	0.11	4.43	< .001***
Group × Test	0.00	0.28	0.00	.997
Group × Dur	0.09	0.27	0.32	.750
Test × Dur	−0.25	0.10	−2.39	.017*
Spec × Group × Test	0.06	0.22	0.29	.772
Group × Test × Dur	−0.60	0.20	−2.97	.003**

Table 4.

Statistical output of the mixed-effects logistic regression model for the experimental and control groups.

Group	Fixed Effect	Estimate (β)	Standard Error	*z*	*p*
Experimental	(Intercept)	0.90	0.17	5.36	< .001***
	Spec	0.68	0.14	4.76	< .001***
	Test	−0.04	0.16	−0.23	.819
	Dur	1.22	0.18	6.69	< .001***
	Spec × Test	0.53	0.13	3.96	< .001***
	Test × Dur	−0.61	0.15	−4.07	< .001***
Control	(Intercept)	0.42	0.13	3.24	.001**
	Spec	0.71	0.21	3.32	.001**
	Test	0.02	0.25	0.06	.949
	Dur	1.13	0.21	5.49	< .001***
	Spec × Test	0.47	0.18	2.57	.010*
	Test × Dur	0.17	0.16	1.04	.298