# How does having a good ear and memory matter for successful second language phonological learning and teaching? An experimental study

by Yaoyao Ruan

British Council's Master's Dissertation Awards 2022

Commendation

# How does having a good ear and memory matter for successful second language phonological learning and teaching? An experimental study.

Yaoyao Ruan

MA Teaching of English to Speakers of Other Langua
Department of Culture, Communication & Media, Institute of Education, UCL

September 2021

# Acknowledgements

I would like to express my appreciation to everyone who assisted me throughout the project.

I would first like to thank my supervisor, Dr Kazuya Saito. Without the guidance, support and encouragement of Kazuya, it would be impossible to carry out such a project. Kazuya's guidance was invaluable, not only in shaping the design of the project and this paper, but also in teaching me how to think and act like a researcher；his passion and dedication have developed my interest in L2 speech research.

Furthermore, my thanks go to two members of Dr Kazuya Saito's team, Katya Petrova and Magdalena Kachlicka. Without their patient explanation and technical support for the Gorilla tasks, this project would not have been possible. I would also like to show my gratitude to all of the students at Wenzhou Medical University who participated in this project with great patience and care. My special thanks go to my colleague, Beibei Zhao, for her great contribution to the recruitment and coordination. I am equally indebted to Philippa Valley and Nathan Thomas, who generously offered me their time and insightful advice when things did not go smoothly. I would also like to acknowledge the raters (evaluators) who dedicated their time and took great care of the evaluation.

Finally, I dedicate this paper to my parents, who have sacrificed their priorities and made considerable efforts to support my study. They are my closet friends and my strongest support.

## Abstract

The purpose of this study was to investigate the acquisitional value of form-focused instruction (FFI) with recasts in second language speech development: the perception and production of the English tense vowel [i] and the lax vowel [ɪ] by adult Chinese learners. As language aptitude is a fundamental property for L2 learning, this study also examined whether the effectiveness of FFI and recasts was influenced by different aspects of auditory processing and the phonological working memory. The participants of the study were 55 college-level adult learners of English as a foreign language (EFL), who were randomly assigned to the experimental group and the control group. Both groups received a 1.5-hour meaning-oriented English session. A pre- and post-test design was employed to detect any improvement in the learners' perception and production of English [i] and [ɪ]. Participants' aptitude (auditory processing and phonological working memory) was measured via three psychoacoustic discrimination tests and two visual and text-entry digit span tests, respectively. Statistical comparisons yield three main findings: (a) the experimental group significantly outperformed the control group on perception learning gains, and instructional gains were less significant in production than in perception; (b) individual differences in auditory processing and phonological working memory partly explained the effectiveness of FFI and recasts in different aspects of second language segmental pronunciation learning; (c) auditory processing and phonological working memory are two dissociable aptitudes and represent different perceptual-cognitive abilities.

***Keywords:*** *recast, foreign language aptitude, auditory processing, phonological memory, second language pronunciation learning*

## Table of Contents

# List of Figures

## List of Tables

# Introduction

Over the past five decades, scholars have examined how instruction can help second language (L2) learners develop their linguistic abilities in the most efficient and effective way. Although numerous empirical studies have shown that explicit instruction (focus on linguistic forms, or *focus on formS*) is more effective than implicit instruction (focusing on meaning, or *focus on form*) regarding linguistic learning gains, the generalizability of linguistic knowledge obtained from decontextualized focus-on-formS instruction to real-life contexts remains questionable (Norris & Ortega, 2000). In recent years, there has been growing support for implicit instruction. Scholars have agreed that drawing learners' attention to linguistic constructions when they engage in meaning-oriented activities in classroom settings (i.e., communicative FFI) can help L2 learners generalize what they have learned from instruction to future communicative settings (Spada & Tomita, 2010). The existing literature has been mainly concerned with the impact of FFI on L2 lexicogrammar learning (see Li, 2010). From about 10 years ago, research began to focus on the amenability of the approach to L2 phonological learning, one remarkable finding of which is that communicative FFI was especially effective in L2 pronunciation learning when combined with implicit corrective feedback such as recasts and prompts (e.g., Saito, 2013; Lee & Lyster, 2016; Gooch et al., 2016). However, empirical evidence on the effectiveness of communicative FFI in L2 speech development is still limited. Thus, this study aim3e to provide more empirical evidence for this topic.

L2 acquisition[1] is highly subject to individual differences in cognitive and perceptual abilities, or *aptitude*. Learners with different cognitive and perceptual abilities achieve various levels of proficiency even when they are exposed to L2 input of similar amount and quality. In instructed L2 learning, *aptitude-treatment interaction* (ATI) research is the type of research that investigate how aptitude factors mediate the effectiveness of instructional approaches (Skehan, 2016). However, the majority of ATI studies to date has exclusively focused on L2 lexicogrammar learning and explicit instruction. L2 speech development and implicit instruction are marginalized topics within this type of research. Therefore, this study set out to investigate the mediating effects of auditory processing and phonological working memory, two aptitude factors that are pivotal to L2 speech development, on FFI and oral recasts to address these research gaps.

[1] This paper does not distinguish between L2 and FL acquisition. L2 is used as an umbrella term for both contexts.

This paper reports a quasi-experimental study that examined the acquisitional values of communicative FFI with recasts in Mandarin Chinese learners' acquisition of English [i] and [ɪ]. This is also the first study that examined how auditory processing and phonological working memory uniquely mediated L2 pronunciation learning in an simulated classroom setting.

## Review of Literature

### Instructed L2 Speech Learning

The ultimate goal of learning a language is to use it to communicate. In instructed L2 speech learning, arguably, the biggest challenge is how to transfer the linguistic knowledge and speech skills acquired in the classroom into real-life communication. For language practitioners and material developers, finding pedagogical approaches and designing materials that help learners achieve this classroom-to-reality transfer is an enduring topic. Despite the importance of communicative competence, decontextualized and language-focused explicit instruction (i.e., focus on formS) has remained dominant in L2 classrooms (Saito & Plonsky, 2019). Such dominance may be due to the fact that pronunciation requires both linguistic knowledge (pronunciation rules) and the manipulation of articulatory organs (Saito & Lyster, 2012), thus leaving little room for contextualized and meaning-focused activities in limited class time. While a large number of studies on the effects of formS-focused approach on L2 pronunciation yield positive results at a controlled speech level (e.g., Derwing & Munro, 2005), the effects on spontaneous production have been found rather discouraging (e.g., Elliott, 1997). This calls for instruction that integrates linguistic forms and communicative meaning-oriented tasks (i.e., communicative FFI) in L2 speech learning.

There are two different types of FFI techniques, i.e., (a) the proactive approach (i.e., creating tasks wherein learners are required to use certain linguistic structures accurately with a view of successful task completion) and (b) the reactive approach (i.e., providing corrective feedback in response to the occurrence of linguistic errors). Overall, communicative focus on phonetic form significantly impacts various dimensions of L2 speech learning. Saito (2013) was among the first to compare FFI without corrective feedback and FFI with corrective feedback, and the study revealed that while FFI itself could facilitate perception and controlled and spontaneous production in trained lexical items (items that has been taught in instruction), FFI and recasts could promote learners' generalizability in perception and production, meaning that learners could transfer newly learned phonetic knowledge to novel items out of the classroom. The acquisitional values of FFI and FFI with corrective feedback have been confirmed by many subsequent studies (e.g., Lee & Lyster, 2016 for perception; Parlak & Ziegler, 2016 for controlled and spontaneous production).

Much scholarly attention has also been directed towards examining how the instructional effectiveness can be associated with a range of input variables, such as visual enhancement (Indrarathne & Kormos, 2018), task complexity (Kourtali & Révész, 2019), the length of instruction (Munoz, 2014), and mode of instruction (face-to-face vs. computer-mediated; Parlak & Ziegler, 2016). Emerging studies have been paying special attention to the associations between learner variables (i.e., individual differences) and effects of FFI, such as motivation (Jiang et al., 2016), anxiety (Rassaei, 2015), and language aptitude (e.g., attention and working memory, Indrarathne & Kormos, 2018). Most of these studies are concerned with lexicogrammar, and there is a serious lag in research on individual differences and communicative FFI in L2 speech learning. To the author's best knowledge, no study so far has worked on the mediating effects of auditory processing on FFI, or two different constructs of aptitude factors (i.e., auditory processing and working memory) in post-pubertal L2 speech learning.

## Conceptualization and Measurement of Aptitude and L2 Research

There is ample evidence that even when individuals engage in the same type of instruction for the same amount of time, their L2 outcomes often differ. This is arguably because individuals differ in their perceptual and cognitive abilities to notice, elaborate, and make the most of every input opportunity. Such relevant are generally termed as "aptitude". In Meisel's (2011) model, apart from universal grammar, FL aptitude comprises a domain-specific *language acquisition device*, comprising processing mechanisms such as sound processing, and a domain-general *language making capacity* including working memory, pattern making abilities, etc.

Aptitude is usually measured by aptitude batteries. Carroll and Sapon (1959) pioneered the measurements by developing the Modern Language Aptitude Test (MLAT) battery. MLAT (2012) has five components: number learning (recalling numbers delivered by audio), phonetic scripts (matching speech sounds to phonetic symbols), spelling cues (learning spelling rules from pronunciation), words in sentence (figuring out grammatical rules from key words in sentences), and paired associations (memorizing vocabularies and their meanings in another language). These test components correspond to four kinds of aptitude: associative memory, inductive language learning ability, grammatical sensitivity, and phonetic coding. The MLAT battery inspired ATI research, in which aptitude tests are typically administered before the instruction, and the outcomes of learning are assessed after instruction to reveal the correlations between aptitude measures and achievement. This is categorized as the 'macro' approach of aptitude research by Skehan (2002).

Extensive macro ATI studies using MLAT (e.g., Erlam, 2005; Hwu & Sun, 2012; Robinson, 2002) have provided solid evidence that aptitude is positively associated with L2 learning success, but it failed to specify the role of aptitude in acquisitional *processes*. Thus, the 'micro' approach was created, aiming to examine aptitude effects in different stages of L2 acquisition including:

> "- *input processing*
> - *noticing*
> - *pattern identification*
> - *complexification (extending, restructuring, integrating)*
> - *handling feedback*
> - *error avoidance*
> - *automatization*
> - *creating new repertoire, achieving salience*
> - *lexicalisin*g"
>
> (Skehan, 2016, p.18)

The micro approach encouraged the exploration of the nature of aptitude (explicit vs. implicit) and the creation of subsequent testing batteries, such as the LLAMA Language Aptitude Tests (Meara, 2005). LLAMA consists of four subtests: LLAMA B, E and F are tests of vocabulary learning, sound-symbol association and grammatical inferencing, and LLAMA D is a test of sound recognition (Meara, 2005). It was argued that the abilities to learn novel words, associate sounds to symbols and infer grammatical rules were associated with one another and tapped into the explicit cognitive processing, while recognising sounds was an implicit dimension of cognition as there was no explicit representation (Granena, 2012 & 2013). This argument was largely supported by investigations into impact of aptitude on *handling feedback* (e.g., Yilmaz, 2013; Yilmaz & Granena, 2015, 2021): explicit language aptitude (indexed by LLAMA B, E and F) only predicted after-treatment outcomes under explicit feedback conditions (explicit correction, metalinguistic feedback), while implicit language aptitude (indexed by LLAMA D and Hi-LAB [2] measures of memory and sequencing) had more interactions with outcome measures under implicit corrective feedback conditions (e.g., recast, elicitation). These findings suggest that the explicitness of feedback strategies or pedagogical practices matches the nature of the aptitude constructs that they tap into. Based on findings from research on L2 grammar learning, it is hypothesized that while both auditory processing and phonological working memory would mediate the instructional gains in L2 vowel sounds from FFI and recasts, the impact of auditory processing would be more significant than phonological working memory, as both instructional input and outcome measures are sound-based, which are considered implicit.

## Phonological Working Memory and L2 Speech Learning

This study is primarily conceptualized within the theoretical framework of Baddeley's *multi-component model* of working memory (WM), initially proposed by Baddeley and Hitch (1974). In Baddeley's revised model, WM is composed of a supervisory control system known as the *central executive* (CE), and three slave systems—*phonological loop*, *visuospatial sketchpad*, and *episodic buffer* (Baddeley, 2000a). CE is responsible for attentional control; phonological loop and visuospatial sketchpad are responsible for maintaining and processing sound-based and visuospatial information respectively, and the episodic buffer acts as an interface between the other two slave systems and long-term memory (LTM) (Baddeley, 2000a). The functioning of the four WM components is subject to their limited capacities and can be measured by certain behavioural tasks. According to Baddeley (2003), the phonological loop is assumed to have two subsystems: (a) a temporary phonological storage, also known as phonological short-term memory (PSTM), and (b) a processing system, the articulatory rehearsal. The former holds phonological information for a few seconds, allegedly two seconds (Baddeley, 2000b), and the latter refreshes the stored information to prevent decay for production. A person's phonological WM capacity (PSTM and complex processing) represents how well his/her phonological loop functions. PSTM is typically assessed with single-component tasks requiring subjects to recall a sequence of items (e.g., digits, words, consonants, non-word) in the presented order, and complex processing is measured with complex tasks, such as the backward digit span tasks, wherein participants are asked to recall sequences of digits in a reverse order (Kaushanskaya & Yoo, 2011).

[2] Hi-LAB was designed to provide a list of predictors of high-level L2 proficiency by Linck et al. (2013).

Robust evidence supports that language processing demands on phonological WM at least to some extent. A considerable amount of literature has reported positive correlations between phonological WM and L2 vocabulary and grammar acquisition (e.g., Atkins & Baddeley, 1998; Ellis & Sinclair, 1996; French & O'Brien, 2008; Masoura & Gathercole, 2005). Not much attention was paid to the relationship between phonological WM and L2 speech learning until the innovative and well-designed study of O'Brien et al. (2007), which found a strong correlation between PSTM capacity and L2 oral fluency development among adult learners. Since then, phonological WM has been shown to be relevant to different aspects of L2 speech development, such as complexity (e.g., Granena & Yilmaz, 2019), phonological processing and perception of L2 sounds (e.g., Darcy et al., 2015), and pronunciation (e.g., Baills et al., 2021). Regarding ATI research, phonological WM does not seem to have been studies specifically in the context of FFI. However, many studies on L2 grammar acquisition demonstrated that individual differences in phonological WM predict learning gains from recasts (Goo, 2012; Révész, 2012) and task-based instruction (Kourtali & Révész, 2019), which is essentially form-focused. Although extensive research has been carried out on WM and L2 acquisition, no single study exists which looks into the effects of phonological WM on L2 speech learning in communicative FFI.

## Auditory Processing and L2 Speech Learning

Speech is one of the most cognitively sophisticated ways in which the human makes use of sounds, as the human is capable of creating infinite permutations and giving them meaning with a limited number of auditory elements such as phonemes and tones (Zatorre et al., 2002). Processing sounds relies on the domain-general neural mechanisms of the human auditory nerve system. This domain-general perceptual ability to represent the spectral and temporal characteristics of sounds is *auditory processing* (AP). It was not until recently that a comprehensive framework of AP came out—the tripartite framework by Saito and colleagues' (2020a), according to which AP consists of *explicit acuity*, *pre-conscious encoding*, and *temporal reproduction*. The first component refers to the accuracy of temporal and spectral discrimination; the second is pre-attentive processing of the frequency of formats; and the last element is the fine motor skills of sequencing and timing sound reproduction. It was also not until recently that sophisticated AP tests were developed. Psychoacoustic thresholds (discrimination tasks of formant, pitch, duration etc.), melodic and rhythmic reproduction and electrophysiological response (also known as frequency following response, FFR) are three of the most commonly used measures of AP abilities (Sun et al., 2021).

AP abilities were primarily measured to diagnose auditory dysfunction, and deficits in AP is likely to cause slower language development (e.g., Kalashnikova et al., 2019). Studies have demonstrated that patients with AP disorders tended to have difficulties reading (Javitt & Sweet, 2015), listening (Zhang et al., 2012), recognising and producing speech sounds (McKinney et al., 2017) in their first language (L1). To test the hypothesis that AP also serves as the threshold of L2 acquisition, certain scholars have examined the generalizability of the topic in the context of post-pubertal L2 speech learning. For example, Kachlicka et al. (2019)'s study revealed that psychoacoustic thresholds were better predictors of English vowel perception than experience-related factors such as age of acquisition (AOA), length of residence (LOR), and length of instructional training. This finding was replicated by Saito et al. (2020a). Their large-scale investigation discovered that while the auditory precision was determinant of L2 proficiency, its predictive power was relatively weaker in inexperienced learners (LOR < 4 months) than that in experienced learners (LOR > 6 years). Longitudinal investigations have demonstrated that those with higher levels of auditory processing tend to yield more considerable improvement in pronunciation proficiency when they engage in immersion experience (e.g., Sun et al., 2021).

In contrast to the large volume of published studies describing the role of AP in naturalistic L2 acquisition, not much empirical evidence can be found in existing literature on the role of AP in classroom settings. The most recent and relevant study is Saito et al., (2021), which found a moderate-to-strong correlation between audio-motor integration scores (i.e., the rhythmic and melodic reproduction ability) and lexicogrammar accuracy among 39 Vietnamese adult learners in a classroom setting; explicit acuity, however, did not significantly impact learners' performance (fluency or accuracy) in spontaneous production tasks. Although there is some evidence that AP mediates the impact of explicit phonetic training (e.g., Chandrasekaran et al., 2010 for the intensive exposure to target sounds), no empirical studies have ever explored how those with varied AP abilities can differentially benefit from more implicit communicative FFI.

## The Current Study

### Motivation

Although evidence from research have proven the acquisitional value of communicative FFI on L2 pronunciation development in adulthood, such as in Japanese speakers learning English /ɹ/ (Saito, 2013), and Korean speakers learning English /ɹ/ (Gooch et al., 2016) and English [i] and [ɪ] (Lee & Lyster, 2016), little evidence was provided by investigations into the learning of English pronunciation by native Mandarin learners, who struggle with English vowel acquisition due to the relatively smaller vowel inventory of Mandarin than English.

Another issue was that while there is ample evidence for the fundamental role of AP in post-pubertal L2 speech development in a naturalistic context, research to date has not yet determined the role of AP in an instructed context. There is some methodological discussion that the way AP is measured via behavioural tasks (e.g., discrimination tasks; for details, see the Method section) inevitably taps into a range of cognitive abilities, such as attentional control and memory (Snowling et al., 2018). As reviewed earlier, WM is also found to be instrumental to L2 speech perception and production. It is important to examine how AP and WM *uniquely* influence the outcomes of L2 speech learning within the same research design. Furthermore, there is a limited amount of empirical evidence for the relative weights of AP and WM in instructed L2 *speech* learning. Therefore, the current study set out to address the following research questions:

1. To what extent does communicative FFI  can help Chinese learners of English improve their L2 vowel acquisition (English [i] and [ɪ])?
2. To what extent do individual differences in AP and phonological WM mediate the learning outcomes?
3. What is the relationship between the two domain-general aptitude, i.e., AP and phonological WM?

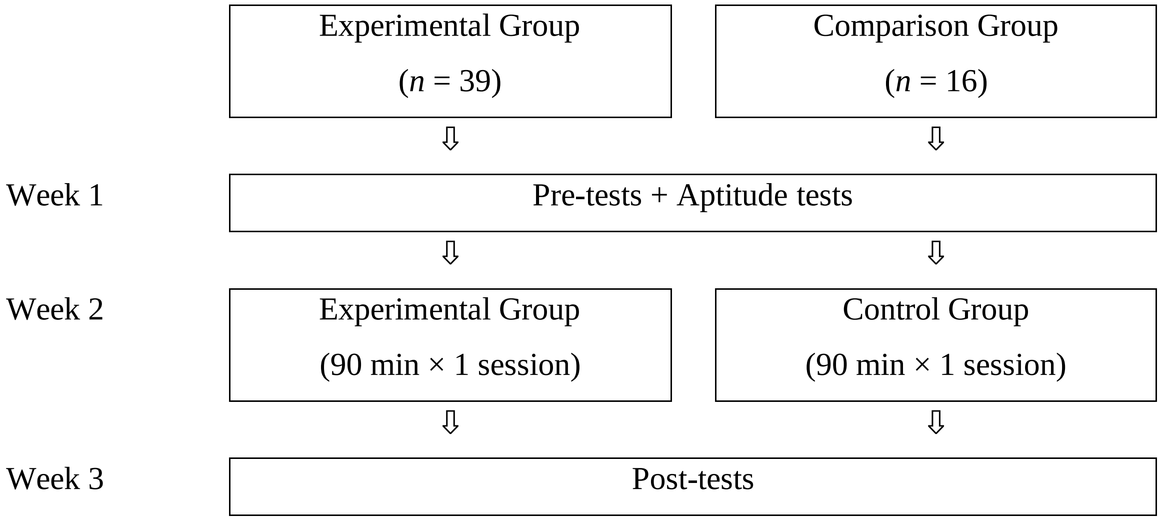
## Method

### Design

This study employed a quasi-experimental pre- and post-test design, with 55 participants randomly assigned to the experimental group (*n* = 39) or the control group (*n* = 16). Before the experiment, both groups took the pre-test and aptitude tests (AP and phonological WM tasks). Following the pre-test, both the experimental and the control groups received a one-and-half-hour instruction on English argumentative skills. The instruction for the experimental group included FFI with recasts on their pronunciation of English [i] and [ɪ], while the instruction for the control group did not have any emphasis on pronunciation or any target words used in the experimental group. Three measures were adopted to assess participants' pre- and post-instructional performance on English [i] and [ɪ]—perception, controlled production, and spontaneous production tests, and the outcomes were analysed in two different lexical conditions (i.e., trained and untrained), as described in Saito (2013). Finally, the accuracy of the English [i] and [ɪ] production data produced in the pre- and post-tests was evaluated by four native-speaking (NS) listeners in the UK. The design of the study was visually summarized in Figure 1.

**Figure 1**

*Summary of Research Design*



## Participants

### *Learners*

The learner participants of this study were young adult Chinese learners of English from a university in China. None of the participants reported hearing impairments. During the participant recruitment phase, ads were distributed to the non-language and non-linguistics major undergraduate students from the university online by the researcher or face-to-face by their English teacher. Interested participants contacted the researcher to sign the consent form and set up a date for the pre-test. A total of 58 volunteers initially participated in the current study, and three of them withdrew from the project for personal reasons.

The mean age of the whole sample was 19.1 years old, ranging from 18 to 22.  Forty-three learners had taken the IELTS test in the past two years, with an average overall score 6.3, ranging from 5.5 to 7, and an average score of 5.7 for speaking. According to the descriptor of CEFR levels and multiple English tests by Cambridge Assessment English, the English proficiency of the participants in the current study belonged to CEFR B2, which is equivalent to IELTS 5.5-6.5 (Cambridge University Press & Assessment, n.d.) The other 12 learners have passed other English tests (e.g., TOEFL, CET-4, or English tests by the university) with an equivalence of CEFR B2. Eleven participants reported immersion experience in English-speaking countries (Australia, Canada, New Zealand, the UK and the US) for less than one month. Table 1 summarizes the demographic information of the learners by treatment condition. After the pre-test, the researcher randomly assigned 16 participants to the control group and the rest to the experimental group ($n$ = 39).

**Table 1**

*Participant Information by Group*

| | Experimental ($n$ = 39) | Control ($n$ = 16) |
|---|---|---|
| Age | $M$ = 19.2, $SD$ = 1.00 | $M$ = 18.8, $SD$ = .58 |
| Gender | 22 females, 17 males | 6 females, 10 males |
| IELTS speaking result | $M$ = 5.8, $SD$ = .32 ($n$ = 27) [a] | $M$ = 5.7, $SD$ = .93 |
| No immersion experience | 32 learners | 12 learners |
| Immersion experience $0 < x \leq 1$month | 7 learners | 4 learners |

[a] Of the 39 participants in the experimental group, 27 learners have taken the IELTS test, and all others have passed equivalent English tests (e.g., CET-4, English tests by the university).

*Instructor*

The instructor (the investigator) for both groups is a native Mandarin speaker with advanced L2 English proficiency. She had been teaching English as an FL to Mandarin Chinese speakers for about five years and was studying for her MA TESOL degree at the time of this project. The instructor prepared the content of the session based on the materials developed by Saito (2013), and she practiced the lessons for the two groups under the guidance of her supervisor before the project started.

*Listeners*

To assess the learners' performance in the production tests, the researcher recruited NS of English in online communities of her university. Four native English speakers (3 females, 1 male) were selected based on their language background and experiences. While the male rater was a native American English speaker, who has been studying for his master's degree and PhD in the United Kingdom, the other three raters were NS of Southern British English, also known as the Standard English. All the native-speaking listeners were experienced in ESL/EFL teaching or in linguistics-related domains. The recruitment of expert raters fits the objective of the current study—the development of L2 segmental accuracy, as professional listeners tend to rely on phonological information for their assessment (Saito, 2021). To try to avoid rater bias caused by listeners' accent familiarity/unfamiliarity and L2 backgrounds (Winke et al., 2013), two listeners were naïve raters, and the other two listeners were experienced raters. The naïve raters reported being unfamiliar with Chinese-accented English (*M* = 2.5 on a 6-point Likert scale, 1 = not at all, 6 = very much) and having infrequent contact with Chinese learners of English (*M* = 3 on a 6-point Likert scale, 1 = very infrequent, 6 = very often), and neither of them had Chinese learning experience. By contrast, the experienced raters reported being familiar with Chinese-accented English (*M* = 5.5) and having frequent contact with Chinese learners of English (*M* = 6) with one of them having learned and taught Chinese as a FL.

## FFI and Recasts

*Content of Instruction*

Most previous studies have reported that there is a stronger relationship between measures of language aptitude and explicit instruction and corrective feedback strategies than implicit ones in intensive learning (Skehan, 2015). Therefore, the current study adopted FFI incorporating metalinguistic information and relatively explicit recasts (i.e., partial recasts) as described in Saito (2013) to induce the best possible reaction between aptitude and pedagogical strategies in the short teaching time (90 minutes). FFI and recasts were embedded in meaning-oriented lesson on English argumentation skills, which aimed to train students to develop critical thinking, express opinions and counterarguments, and provide adequate justifications. The instruction took the form of an in-class debate, which has proved effective for the development of oral argumentation skills in L2 classrooms (Majidi et al., 2021). The 26 minimally paired words with [i] and [ɪ] sounds (see "Trained Items" in Table 2) were embedded in the topics that were discussed and debated in class (see Appendix A for teaching materials).

All the topics were topical phenomena or issues familiar to Chinese university students. The 90-minute session for the control group also concentrated on English argumentation skills, but all the target words were carefully avoided, and no FFI or recasts were provided.

### *Target Sounds: English [i] and [ɪ]*

The tense vowel [i] (as in "seat") and the lax vowel [ɪ] (as in "sit") were the target sounds in the current study, and they differ in terms of quantity (length) and quality (articulatory features) (Hillenbrand & Clark, 2000). Extensive studies have found that L1 Mandarin learners of English have difficulties distinguishing between [i] and [ɪ] in English and tend to categorize both vowels into [i] (e.g., C. Wang, 1988; X. Wang, 1997, 2002, 2006; Wang & Munro, 1999). These findings are in accordance with the Perceptual Assimilation Model (PAM, Best, 1995) and the Speech Learning Model (SLM, Flege, 1995). According to these two models, the accuracy of discrimination depends on how learners assimilate L2 sounds into their L2 phonetic categories, and the more significantly a L2 sound differs from the most similar L1 sound, the easier it is to lead to 'Two Category Assimilation' (Best, 1995) or to establish a new phonetic category (Flege, 1995), which facilitates near-native discrimination. Given that only the sound [i] exists in the Mandarin vowel inventory, L1 Mandarin learners of English are likely to perceive English [i] and [ɪ] as equivalences of the Mandarin [i] (Single Category Assimilation, Best, 1995), or to classify English [i] and [ɪ] as the Mandarin [i] while feeling that the English [i] is closer to the Mandarin [i] than the English [ɪ] (Category Goodness Assimilation, Best, 1995). Either Single Category or Category Goodness prevents native Mandarin learners from forming new phonetic categories for the English [i] and [ɪ], thus hindering native-like perception and production. Moreover, there are no contrasts of tense and lax vowels in Mandarin (Smith et al, 2019), for which native Mandarin speakers tend to depend significantly on the duration of the vowel sounds to distinguish between them, ignoring the variations in quality (Wang & Munro, 1999). In summary, Mandarin learners of English may require more time and assistance (e.g., explicit instruction and explicit corrective feedback) to master the phonemic contrast between the English [i] and [ɪ].

**Table 2**

*Fifty Tokens in the Proficiency tests*

| Trained Items | Untrained Items | Distracters [a] |
|---|---|---|
| 1. bean-bin | 1. feet-fit | 1. bag-beg |
| 2. beat-bit | 2. keys-kiss | 2. man-men |
| 3. cheap-chip | 3. lead-lid | 3. mass-mess |
| 4. feel-fill | 4. least-list | 4. pat-pet |
| 5. heel-hill | 5. meal-mill | |
| 6. heat-hit | 6. peak-pick | |
| 7. leave-live | 7. sleep-slip | |
| 8. reach-rich | 8. steel-still | |
| 9. read-rid | | |
| 10. scene-sin | | |
| 11. seat-sit | | |
| 12. seek-sick | | |
| 13. sheep-ship | | |

*Note*. The lexical items are listed alphabetically in the table. They were presented in random order in the tests.

[a] The controlled production test was the only test involving distracters.

*FFI and Recast Treatment*

The results of the eye-tracking study conducted by Indrarathne and Kormos (2018) showed that language aptitude might be a stronger predictor for L2 learning results under explicit instructional conditions than implicit conditions, since their attention is explicitly directed to the target linguistic feature. To direct learners' attention to [i] and [ɪ] in an essentially implicit instruction (i.e., FFI) , the current study follows Saito (2013) and Lee and Lyster (2016) by incorporating the following strategies in FFI: (a) metalinguistic information—the instructors provided exaggerated pronunciation of the target sounds and explained the articulation configurations (i.e., standard positions of a speaker' articulatory organs), (b) input enhancement—typographical enhancement was added to visual input in the teaching materials (i.e., underlining and highlighting the target words), and (c) awareness tasks—learners were encouraged to compare [i] and [ɪ] through warm-up games at the start of the lesson (see Appendix B).

Learners in the experimental group were also consistently provided recasts by the instructor for their unclear, erroneous or unintelligible pronunciations of the target words. To increase the perceived prominence of the oral corrective feedback, recasts were constantly supplied for the mispronounced word rather than the entire phrase (i.e., partial recasts rather than full recasts) in a falling tone (Sheen, 2011). The instructor gave an opportunity for learners to self-correct by leaving a two-second margin following each recast. Only one recast was given to each mispronunciation, unless the student failed to hear it and requested to repeat it. This means that even if learners responded to a recast with a reaction other than repair (i.e., need repair or no repair), no more recasts or other remedial hints were provided. See the following examples that demonstrated three situations of recasts (i.e., repair, need repair and no repair) extracted from one of the sessions, according to Lyster and Ranta's (1997) scheme of learner uptake (learners' responses to corrective feedback).

*Repair*
S: People shouldn't seek doctor's help immediately when they feel seek* [sick].
T: Sick.  (RECAST)
S: Sick. (REPAIR)
(Recast stops.)

*Need repair*
S: Young people should always give up their seats to the elderly when they have no place to seat* [sit].
T: Sit.  (RECAST)
S: Seat* [sit]. (NEED REPAIR)
(Recast stops.)

*No repair*
S: I think reading books is the best way to get read* [rid] of stupidity.
T: Rid.  (RECAST)
S: Because there is a lot of knowledge in books. (NO REPAIR)
(Recast stops.)

## Test Instruments

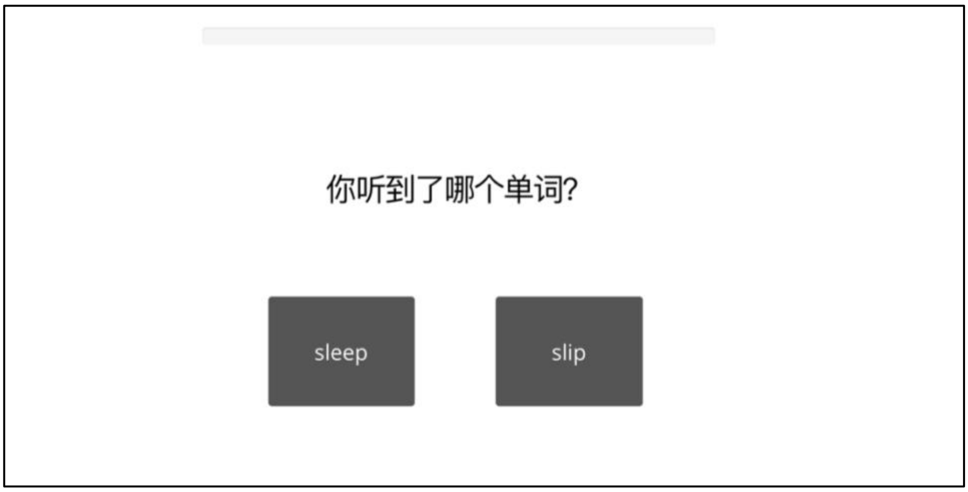
### *Proficiency tests*

To examine the treatment effects of FFI and recasts on learners' perceptual accuracy and productive competence of [i] and [ɪ], (a) a perception test, (b) a controlled production test, and (c) a spontaneous production test were employed, as described in Saito (2013). To prevent test takers from excessively focusing on form, the pre- and post-tests were performed in the order of (c), (b), (a). Learners accessed the tests via a web-based application *Gorilla* (Anwyl-Irvine et al., 2020) due to COVID-19 restrictions.

The testing materials included both trained (i.e., words that occurred in the experimental materials) and untrained (i.e., words that learners did not encounter in the experimental materials) items. The untrained words served as generalizability words, indicating if the results of teaching can be applied to novel words. All of the lexical items (*n* = 50) in testing materials are Consonant-Vowel-Consonant (CVC) singletons (see Table 2). They all fall among the first 4,000 most common word families of the British National Corpus, according to vocabulary profiling via Lextutor (Cobb, 2012). In an investigation into English vocabulary size across CEFR levels, Milton (2010) stated that around 3,000 of the most frequently occurring words might be required to just go beyond the basic levels (i.e., A1 and A2), and advanced CEFR levels (i.e., C1 and C2) were related with full recognition of the 5,000 most frequently used words. Thus, for intermediate-level learners (CEFR B1 and B2) in the current study, the impacts of lexical frequency and familiarity on test performance was reduced.

**Figure 2**

*A Screenshot of the Perception Test*

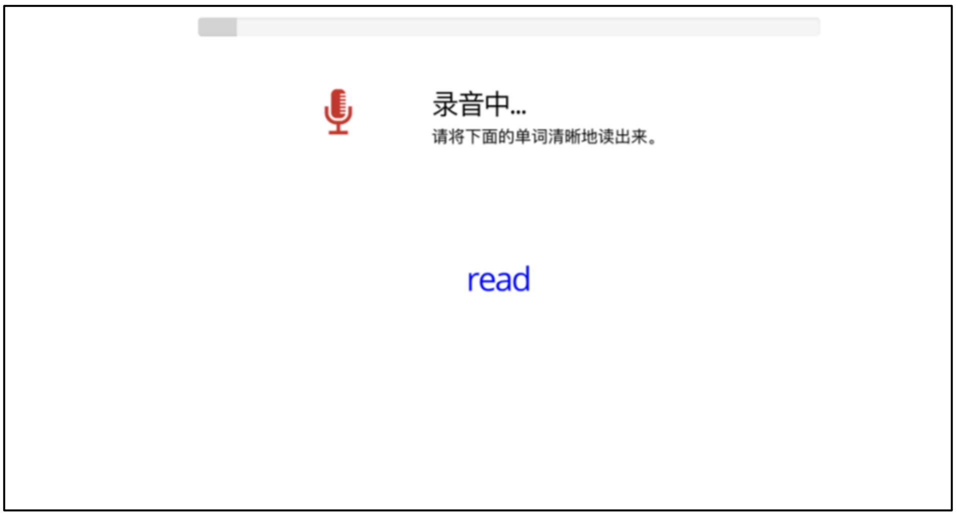


你听到了哪个单词?

**Perception test**. A two-alternative forced identification task was used to assess learners' receptive knowledge of [i] and [ɪ] before the project: participants listened to 42 randomized minimally paired words with [i] and [ɪ] (e.g., "seat-sit") (trained and untrained items in Table 2) via their own headsets, and they were required to identify the word heard by clicking one of the two orthographic options presented on the computer screen. In Figure 2, for example, participants were asked to distinguish “sleep” from “slip” by answering the question “你听到了哪个单词？”(Which word did you hear?). All speech samples were recorded by one male and one female professional voice-over artists, who were native speakers of Standard British English, in isolated studios with professional recording equipment. Each sample was digitalized at a 44,100 Hz sampling rate and normalized.

**Figure 3**

*A Screenshot of the Controlled Production Test*



录音中...

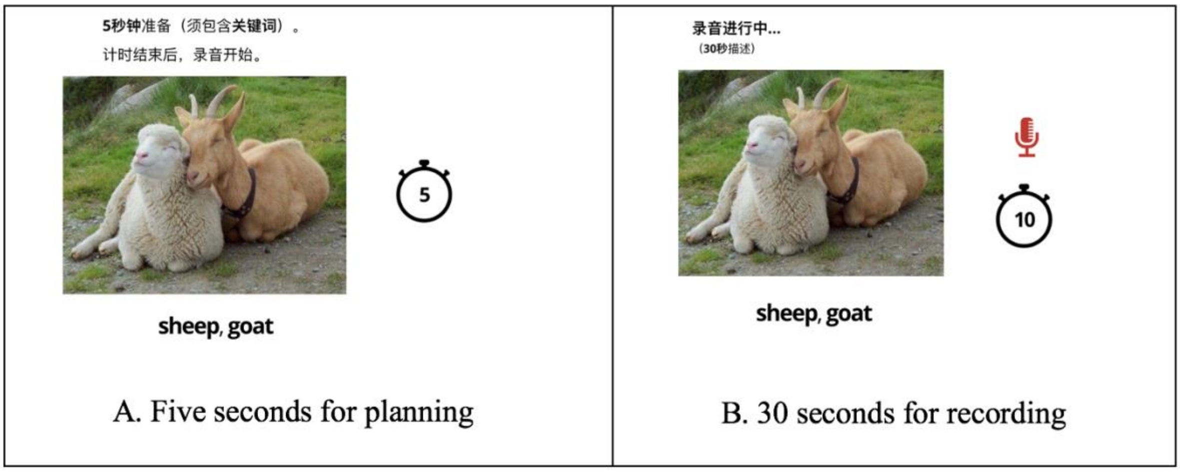
请将下面的单词清晰地读出来。

read

**Production tests**. The production test consisted of a controlled word reading task and a spontaneous picture description task. The controlled production task elicited learners’ production of [i] and [ɪ] by asking them to read a list of 16 words (4 trained, 4 untrained and 8 distracters) (see Table 3). For example, Figure 3 illustrates that participants were asked to read aloud the target word —“read”—clearly (“请将下面的单词清晰地读出来”). Learners’ responses were automatically recorded and saved by Gorilla to the online database, which only the researcher had access to. When the learner made multiple attempts for the same word, the last and complete response was taken as the final answer. The spontaneous task measured learners’ ability to use the target words under time pressure. Learners were given five seconds to look at a picture and read the two key words (one target word, one content-related word) below it on the computer screen. They were then instructed to describe the picture using the two key words within 30 seconds. Figure 4 shows an example where the singleton “sheep” was the target word, and a timer was displayed on the righthand side of the screen to indicate the remaining time for planning (Figure 4-A) and recording (Figure 4-B). There was no other planning time assigned except for the 5 seconds. The automatic recording started and ended as per the time limit. In total, each learner described 8 pictures, contributing 4 trained words and 4 untrained words (see Table 3). The stimuli in the controlled production task and the pictures in the spontaneous task were presented in a random order.

**Figure 4**

*Screenshots of the Spontaneous Test*



5秒钟准备（须包含关键词）。
计时结束后，录音开始。

A. Five seconds for planning

录音进行中...
（30秒描述）

B. 30 seconds for recording

**Table 3**

*Six-teen Tokens in the Controlled and Spontaneous Production Tests*

| Test | Trained Items | Untrained Items |
|---|---|---|
| Controlled Production | read-rid | feet-fit |
| | seek-sick | lead-lid |
| Spontaneous Production | heel-hill | keys-kiss |
| | sheep-ship | peak-pick |

*Listener Judgement*

Four native-speaking raters evaluated participants' pronunciation skills in both controlled and spontaneous production tasks.

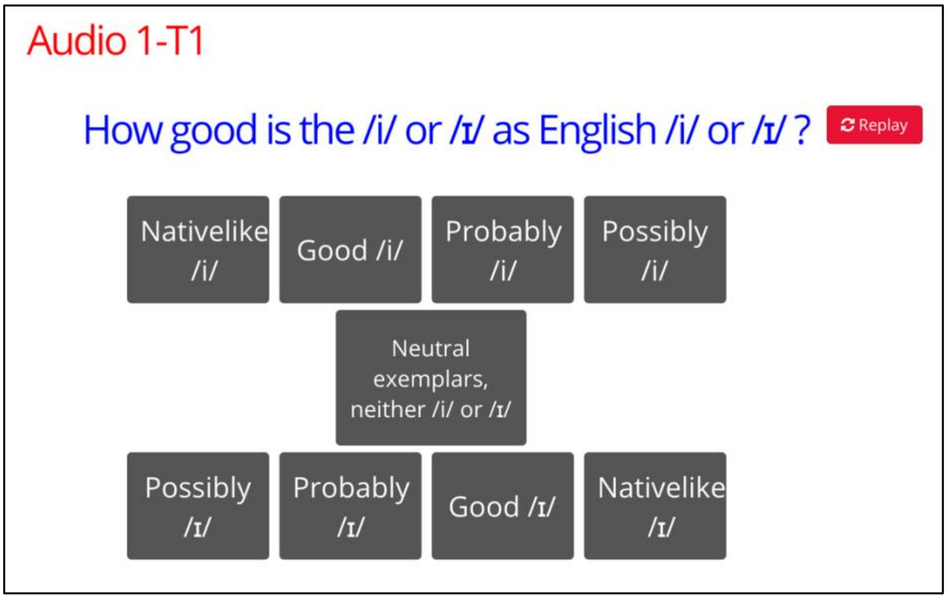
**Material Preparation.** A total of 1,568 speech samples produced by the 49[3] Chinese learners from the pre- and post-tests (49 learners × 8 words × 2 production tasks × 2 tests) were retrieved from Gorilla. Given the workload of the raters, half of the sample ($n$ = 784) was

[3] 6 participants' production data were missing due to various technical issues.

selected for rating (49 learners × 4 words [2 trained + 2 untrained] × 2 production tasks × 2 tests). The researcher selected the first occurrence of a word containing [i] and the first occurrence of a word containing [ɪ] by simple random sampling (see Table 4). Considering the inconsistency and uncontrollability of each learner's device and testing environment, the researcher took care to listen to each audio file multiple times and applied noise reduction where needed. All samples were adjusted to a 44,100 Hz sampling rate and normalized to a -1.0 dB peak amplitude. In the case of words embedded in the spontaneous production, the researcher listened to each sample and trimmed them as much as possible without distorting the sound. All speech tokens were coded and fed into the corresponding blocks ($n$ = 4) on Gorilla with 194 tokens in each block following the order: pre-test controlled, pre-test spontaneous, post-test controlled, and post-test spontaneous.

**Figure 5**

*A Screenshot of the Rating Task*



Audio 1-T1

How good is the /i/ or /ɪ/ as English /i/ or /ɪ/ ?

**Procedure.** For safety reasons due to the global pandemic, the raters accessed the assessment tasks the online platform, Gorilla, using their own laptop. The raters received instructions from the researcher before the evaluation. Listeners assessed the quality of [i] and [ɪ] pronunciation by selecting one of the choices from a 9-point scale descriptor adapted from Flege et al. (1995, as cited in Saito, 2013). The 9-point scale descriptor is as follow:

- 1: Nativelike [i]
- 2: Good [i]

- 3: Probably [i]
- 4: Possibly [i]
- 5: Neutral exemplars, neither [i] nor [ɪ]
- 6: Possibly [ɪ]
- 7: Probably [ɪ]
- 8: Good [ɪ]
- 9: Nativelike [ɪ]

Raters were specifically instructed to score only on the quality of the pronunciation of [i] and [ɪ] but not on other aspects such as accent and segmental accuracy of other consonants or vowels. They were also provided with two example trials to familiarize with the tasks. Practice trials were excluded from the main dataset for statistical analyses.

In the assessment tasks, one token appeared on the screen at a time in a fixed order, and the listeners were allowed to listen to each token up to three times before they made their final judgements. The first time was played automatically, after which the "replay" button became available. Listeners could click on it to replay the audio where needed (see Figure 5). They were also told to note down the code assigned to each token (in the top left corner of Figure 5) when they made judgements they did not intend to and report the wrong decisions to the researcher when the entire evaluation was completed. The full review process (four blocks) was anticipated to take approximately one and a half hours. To reduce rater fatigue, short breaks were provided between blocks.

**Table 4**

*Eight Selected Tokens for Rating in the Controlled and Spontaneous Production Tests*

| Test | Trained Items | Untrained Items |
|---|---|---|
| Controlled Production | read | fit |
| | sick | lead |
| Spontaneous Production | heel | kiss |
| | ship | peak |

*Language Aptitude Test*

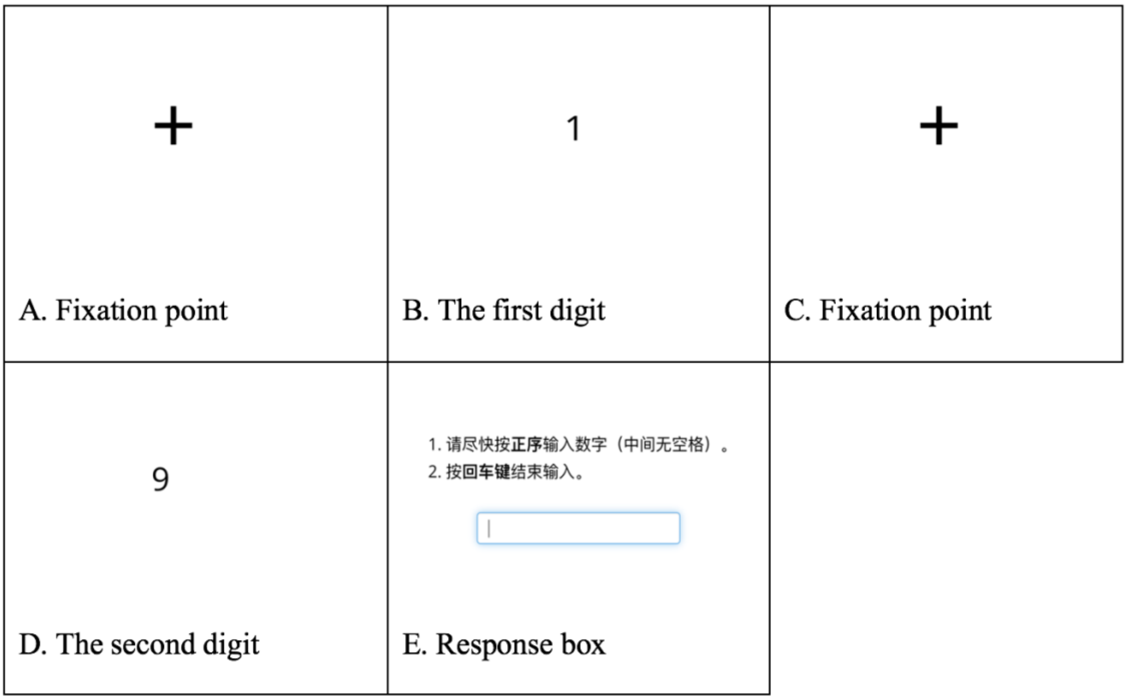
The two kinds of aptitude investigated in this study, verbal WM and AP, were measured by two types of digit span (DS) tasks and three types of psychoacoustic AXB discrimination tasks on Gorilla.

**Measures of Phonological WM.** Phonological WM were measured using the visual and text-entry DS tests adapted and modified from Dean (2020), which originally adapted from Turner & Ridsdale’s (2001, as cited in Dean, 2020) assessment procedure for diagnosing children’s learning difficulties associated with verbal WM. To make the tests more suitable for L1 Mandarin adult learners, the original version was modified in two ways: (a) numbers were presented on the computer screen at the rate of one digit per second rather than the 200ms per digit of the original tests, following the classic practice of DS tests in Wechsler Adult Intelligence Scale IV (WAIS-IV, Wechsler, 2008); (b) in order to avoid the impact of learners' differences in L2 reading comprehension on the test outcomes, the instructions were changed from English to the participants' L1, Mandarin. Pre-recorded sequences (List B of Dean, 2020) were presented to the participants on the screen at the rate of one digit per second with a fixation point in between digits (100ms). Participants were required to recall the numbers by typing in the input box once each sequence finished. Responses were mandatory, and they were required to press “return” on the keyboard to finish responding and advance to the next trial. In Figure 6, the screenshots illustrate the sequence of display of a 2-digit span (A→B→C→D→E).

As a simple span test, the digit span forward test (DSF) was used to define participants’ PSTM capacity, where the complex span task—digit span backward test (DSB)—was the measure for complex WM processing. For the DSF, participants were required to recall the numbers in the order that the numbers were presented. For the DSB, they had to respond by typing numbers in the inverse order as displayed. While there were 9 spans starting at 2 digits up to 10 digits in DSF, there were 8 spans (i.e., 2 digits to 9 digits) in DSB. Each span was composed of two trials, success in at least one of which led to the next level (one digit more than the previous level) until the participants failed to recall either of the two trials of the same length. The capacity of phonological WM was determined by the longest spans that participants entered without error. The two tests were implemented in the order of (a) DSF and (b) DSB, and a break between the tasks was optional. Depending on the exit point and reaction time, this part took about 5—10 minutes to complete.

**Figure 6**

*Screenshots of a Two-digit Sequence*



A. Fixation point

B. The first digit

C. Fixation point

D. The second digit

1. 请尽快按**正序**输入数字（中间无空格）。
2. 按**回车键**结束输入。

E. Response box

**Measures of AP.** The current study used three types of AXB discrimination tests of the AP battery by Kachlicka et al. (2019) to assess the three aspects of explicit acuity—formant, frequency and duration discrimination. The AP battery has been validated and employed by a number of studies investigating the relationship between AP and post-pubertal speech acquisition (e.g., Saito et al., 2020a; Saito et al., 2020b; Sun et al., 2021; Zheng et al., 2020). In each trial, participants listened to three non-verbal sounds and were required to indicate which one (either the first or third) sounded different from the other two by clicking on the number "1" or "3" on the screen with a mouse (see Figure 7). This means that the second sound is always the same as either the first or the third sound in the aspect (i.e., formant, pitch or duration) being tested.

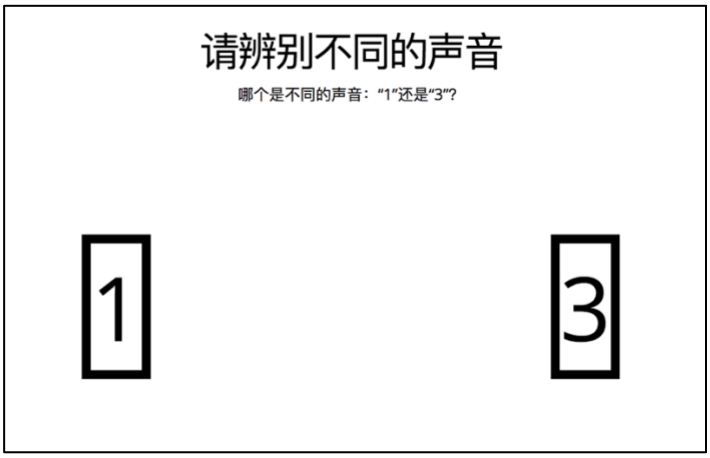
For each test, one reference stimulus and 100 target stimuli were created using the custom MATLAB scripts. Unless otherwise specified, all stimuli were 500-ms-long complex tones of four harmonics with a fundamental frequency (F0) of 330 Hz. For the pitch discrimination test, the reference stimulus was created with F0 at 330 Hz, whereas the target stimuli varied in frequency from 330.3 to 360 Hz with a 0.3-Hz increment. The duration discrimination test ranged with a 2.5 ms step from 252.5 ms to 500 ms. The stimuli for the formant discrimination test were complex tones with F0 at 100 Hz, F1 at 500 Hz, F3 at 2,500

Hz. The reference stimulus was created with the second formant (F2) at 1,500 Hz, whereas the target stimuli had F2 at 1,502–1,700 Hz with an increment of 2 Hz.

Using Levitt’s (1971, as cited in Kachlicka et al., 2019) adaptive up-down procedure, the test began at Level 50 (out of 100 levels) and automatically adjusted its difficulty level in response to the participant's performance: it became 10 steps more difficult when the participant provided three correct responses in a row, or 10 steps easier when the participant gave one incorrect response. The step size decreased to five following the first reversal, from five to two following the second, and from two to one after the third until the task was completed. Each test terminated after 70 trials or 8 reversals (see Kachlicka et al., 2019 for more information).

**Figure 7**

*A Screenshot of AP Discrimination Tasks*



请辨别不同的声音

哪个是不同的声音：“1”还是“3”?

**Statistical Analyses**

The performance of 55 learners was included for the analyses of the perception test. For the controlled and spontaneous production tests, 49 learners’ data were analysed, as the other 6 learners’ responses were not recorded due to technical issues.

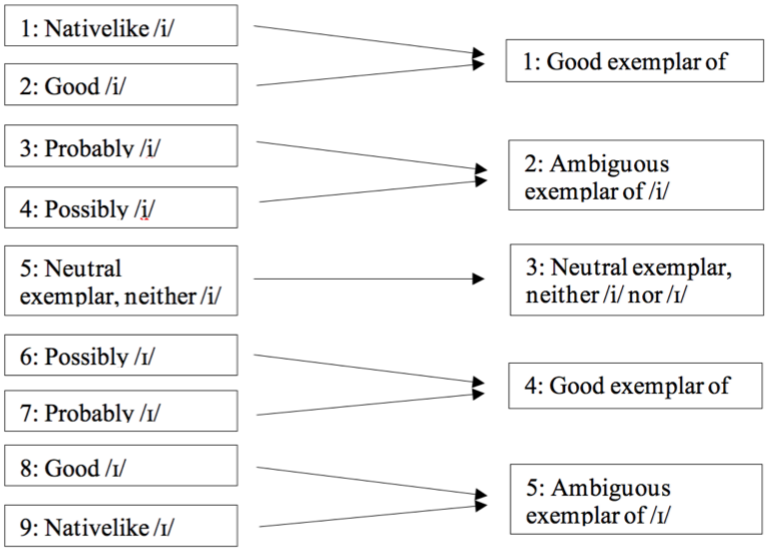
Initial analyses revealed that learners’ progress in the pronunciation of [i] and [ɪ] was not significant on the original 9-point rating scale. In their comprehensive investigation of the relationship between rater experiences, length of rating scales, and judgements of L2 pronunciation, Isaacs and Thomson (2013) reported that no significant group differences were observed in mean scores and inter-rater reliability using the 5- versus 9-point scales. Thus, the rating on original 9-point scale in this study was scaled down to a 5-point scale by collapsing 1 and 2, 3 and 4, 6 and 7, 8 and 9 into one category respectively (see Figure 8). Since the magnitude of the scores was inverted for [i] and [ɪ], the scores for [i] were computed using the formula [6 – n]. For instance, if Participant A was rated *M* = 2 for her pronunciation of [i], her adjusted score would be 4.

For the 9-point rating scale, according to the results of interclass correlation between the four NS raters, Cronbach's alpha was .834 for the entire production data set (*n* = 784), .844 for the controlled production tokens (*n* = 392), and .853 for the spontaneous production tokens (*n* = 392) (*p* = .00). For the 5-point rating scale, Cronbach's alpha was .825 for the entire production data set, .837 for both controlled and spontaneous production tokens (*p* = .00). These results were consistent with those of Isaacs and Thomson (2013): (a) there was a high agreement level among the raters using either the 9- or the 5-point rating scale; (b) judgments based on a 9-point scale was slightly more consistent than judgements based on a 5-point scale, which might be due to "the more restricted scale-step choice" (p.143).

Aptitude variables underwent transformation prior to the analysis. In order to compute learners' spectral processing and temporal processing abilities, the same method described by Kachlicka et al. (2019) was used: the formant and pitch thresholds were transformed to *z*-scores and averaged to create the composite spectral measure, and the *z*-score transformed from the duration threshold represented participants' temporal processing abilities. Their overall AP score is the average of the three z-scores added together and divided by three. Learners' overall phonological WM capacity was computed by the average score of DSF and DSB.

**Figure 8**

*The Transformation from a 9-point Scale to a 5-point Scale*



# Results

## Overall Improvements (Pre- to Post-test)

### *Perception*

The pre- and post-test correct identifications scores (%) were summarized in Table 5 and visually plotted in Figure 9 as per three different conditions: overall ($n$ = 42 items), trained ($n$ = 26 items), and untrained ($n$ = 16 items). Interestingly, there was a clear trend of decreasing performance in the control group (M = 73.9→72.5%), whereas the experimental group's performance increased over time (M = 70.2→74.6%). The results of Kolmogorov-Smirnov test showed that participants' pre-test scores did not significantly differ from normal distribution as to trained conditions ($D$ = .100, $p$ = .603) and untrained conditions ($D$ = .081, $p$ = .835). To find any pre-existing difference in perception accuracy of the target sounds (i.e., English [i] and [ɪ]), participants' total pre-test scores were submitted to a one-way analysis of variance (ANOVA) with Group (Experimental vs. Control) as the between-group factor. No significant Group difference was found at the time of pre-test ($p$ > .05). In order to assess the effects of Time (pre-/post-tests), Group (Experimental vs. Control) and Lexis (trained vs. untrained), repeated-measures ANOVAs were used.

A three-way ANOVA with Group as a between-group factor and Time and Lexis as within-group factors yielded a significant Group × Time interaction effect, $F$ (1, 53) = 6.185, $p$ = .016, $\eta_p^2$ = .105. The analyses of multiple comparisons showed that the experimental group significantly improved their overall scores ($M$ = 70.2 → 74.6%, $p$ = .002, $\eta_p^2$ = .184). In line with Cohen's (1988) benchmarks, the effect size could be considered as medium to large ($\eta_p^2$ = .13-.26); but the control group's performance did not reach statistical significance ($M$ =73.9 → 72.5%, $p$ = .363, $\eta_p^2$ = .016). The three-way Group × Time × Lexis interaction did not reach statistical significance, $F$ (1, 53) = 0.008, $p$ = .927, $\eta_p^2$ = < .001. Interestingly, a significant main effect for Lexis was found, $F$ (1, 53) = 20.883, $p$ < .001, $\eta_p^2$ = .283. According to the pairwise comparisons, learners in both experimental and perception groups performed significantly better on untrained lexical items than trained items at both testing time points (Trained $M$ = 71.2%, $SD$ = .140; Untrained $M$ = 75.0%, $SD$ = .140). Taken together, the results indicate that the experimental group demonstrated a significant improvement in their abilities to identify English [i] and [ɪ] regardless of lexical contexts (trained and untrained).

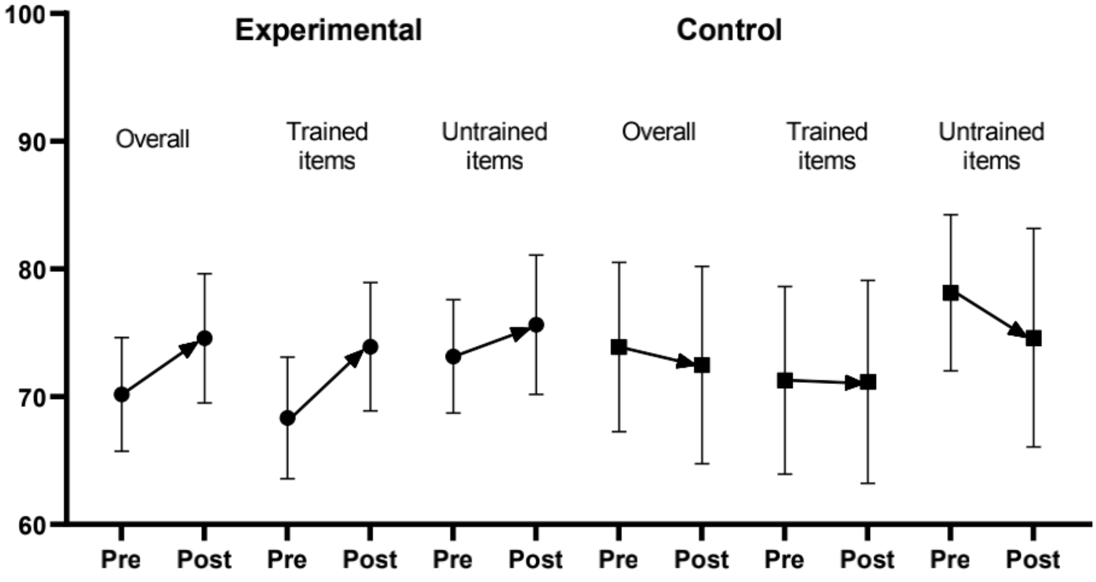
**Table 5**

*Summary of Perception Scores, AP, and WM*

| | Experimental (*n* = 39) | | | | Control (*n* = 16) | | | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | 95% CI | | *M* | *SD* | 95% CI | |
| | | | Lower | Upper | | | Lower | Upper |
| A. L2 speech proficiency (%) | | | | | | | | |
| Overall (pre) | .702 | .14 | .657 | .746 | .739 | .12 | .673 | .805 |
| Trained (pre) | .683 | .15 | .636 | .731 | .713 | .14 | .639 | .786 |
| Untrained (pre) | .732 | .14 | .687 | .776 | .781 | .12 | .720 | .842 |
| Overall (post) | .746 | .16 | .700 | .800 | .725 | .15 | .647 | .802 |
| Trained (post) | .739 | .16 | .689 | .789 | .712 | .15 | .632 | .791 |
| Untrained (post) | .756 | .17 | .702 | .811 | .746 | .16 | .661 | .832 |

**Figure 9**

*95% Confidence Intervals and Mean Values of the Learners' Perception Scores*



*Production*

**Controlled production.** The pre- and post-test rating scores (5-point scale) were summarized in Table 6 and visually plotted in Figure 10 as per three different conditions: overall (*n* = 8 items), trained (*n* = 4 items), and untrained (*n* = 4 items). Evidently, there is an upward trend from the pre-test to the post-test in all cases for both groups, meaning that all participants improved substantially from the pre-test to the post-test in untrained items, particularly the experimental group (M = 3.8→4.6). A Kolmogorov-Smirnov test indicated that participants' pre-test scores did not follow a normal distribution as to the trained (*D* = .190, *p* < .01) and untrained (*D* = .158, *p* < .01) conditions. Therefore, log transformation was applied to transform the skewed data to normality. To find any pre-existing difference in perception accuracy of the target sounds (i.e., English [i] and [ɪ]), participants' total pre-test scores were submitted to a one-way ANOVA with Group (Experimental vs. Control) as the between-group factor. No significant Group difference was found at the time of pre-test (*p* = .421). Similar to the perception data, the results of controlled production in the pre-test and post-test were also analyzed using a three-way ANOVA to examine the effects of Time, Group and Lexis.

**Table 6**

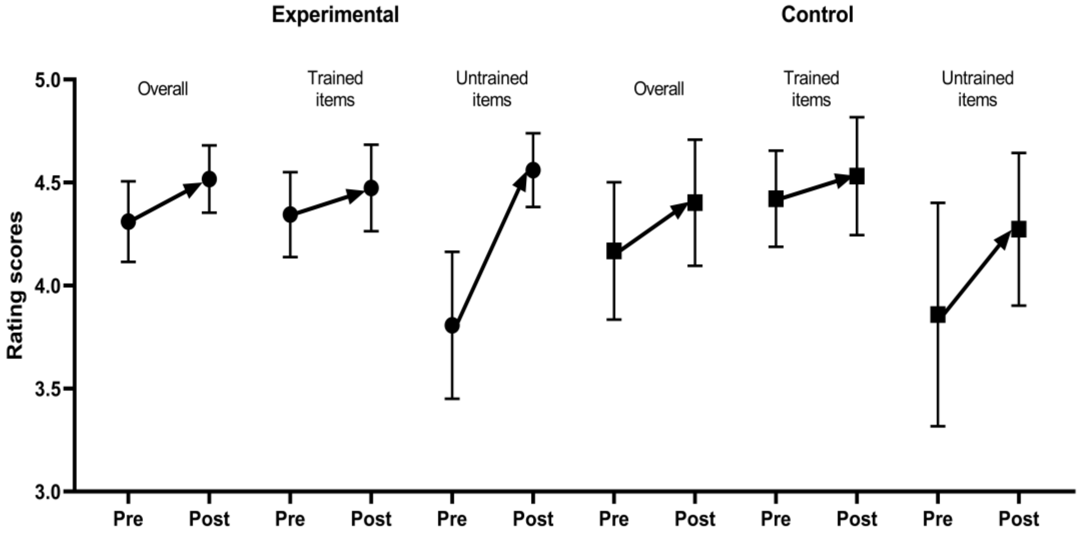
*Summary of Controlled Production Rating Scores*

| | Experimental (*n* = 33) | | | | Control (*n* = 16) | | | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | 95% CI | | *M* | *SD* | 95% CI | |
| | | | Lower | Upper | | | Lower | Upper |
| A. Controlled production proficiency (5 points) | | | | | | | | |
| Overall (pre) | 4.3 | .55 | 4.1 | 4.5 | 4.2 | .63 | 3.8 | 4.5 |
| Trained (pre) | 4.3 | .58 | 4.1 | 4.6 | 4.4 | .44 | 4.2 | 4.7 |
| Untrained (pre) | 3.8 | 1.00 | 3.5 | 4.2 | 3.9 | 1.02 | 3.3 | 4.4 |
| Overall (post) | 4.5 | .46 | 4.4 | 4.7 | 4.4 | .58 | 4.1 | 4.7 |
| Trained (post) | 4.5 | .59 | 4.3 | 4.7 | 4.5 | .54 | 4.2 | 4.8 |
| Untrained (post) | 4.6 | .50 | 4.4 | 4.7 | 4.3 | .70 | 3.9 | 4.6 |

In the three-way ANOVA with Group as between-group factor and Time and Lexis as within-group factors, neither Group × Time nor Group × Time × Lexis were significant (p >.05). Significant main effects were found for Time (*F* (1, 47) = 6.562, *p* = .014), Lexis (*F* (1, 47) = 16.716, *p* < .001), and Time × Lexis (*F* (1, 47) = 10.666, *p* = .002). According to Bonferroni multiple comparisons, both groups made substantial improvements in untrained lexical items (Mean improvement = 0.8 points, *F (1, 47) = 10.796, p = .002*) than trained items (Mean improvement = 0.1 point, *F* (1, 47) = 1.627, *p* = .208). This is likely to stem from the considerable pre-existing differences between the trained (*M* = 4.4, *SD* = .535) and untrained items (*M* = 3.8, *SD* = 1.000). Taken together, the results indicate that (a) the two groups did not differ in their abilities to produce English [i] and [ɪ] in the controlled production task, and (b) less familiar lexical items may allow more room for improvement.

**Figure 10**

*95% Confidence Intervals and Mean Values of the Learners' Controlled Production Rating Scores (1: English [i]—5: English [ɪ])*



**Spontaneous Production.** Table 7 and Figure 11 demonstrate the summary of participants' performance on the spontaneous production tasks in the overall (*n* = 8 items), trained (*n* = 4 items), and untrained (*n* = 4 items) conditions. The period from the pre-test to the post-test witnessed a slight increase (M = 4.2→4.3) in rating scores of learners' pronunciation of English [i] and [ɪ] in the spontaneous production tasks in both groups. Closer inspection of Figure 11 reveals that the experimental group showed a considerable improvement in pronouncing trained items from the pre-test to the post-test (M = 4.0 →4.3), whereas a small but apparent decrease over time was observed from their performance on untrained items (M = 4.4 →4.3). It is interesting that the control group exhibited a notable increase in rating scores from the pre-test to the post-test in untrained items (M = 4.3 →4.5), while their performance on trained items remained stable over time (M = 4.2).

**Table 7**

*Summary of Spontaneous Production Rating Scores*

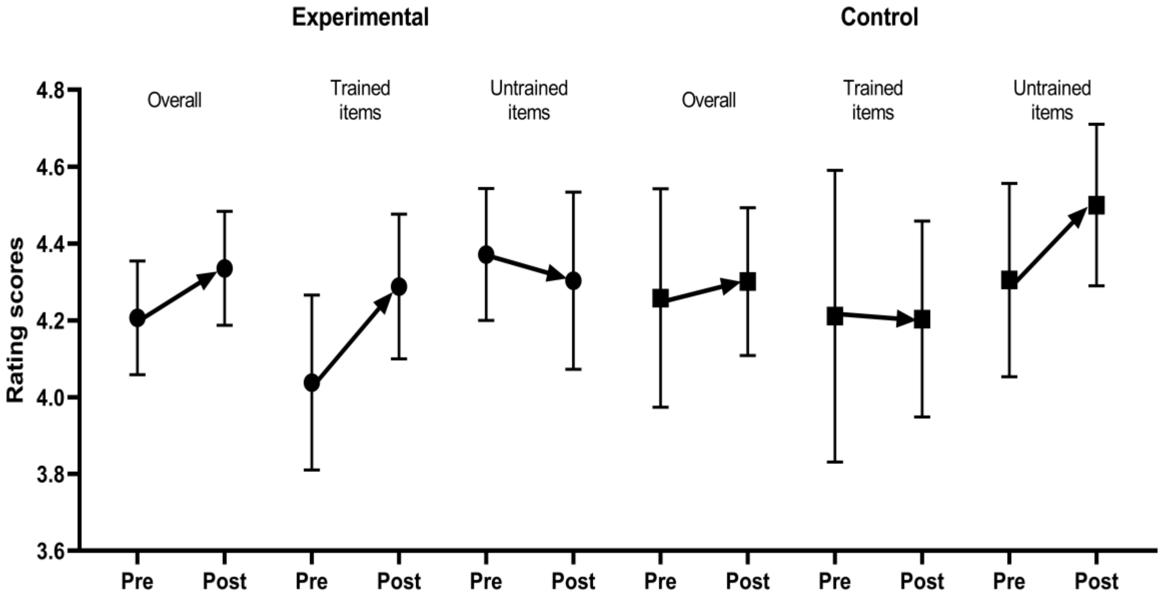
| | Experimental (*n* = 33) | | | | Control (*n* = 16) | | | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | 95% CI | | *M* | *SD* | 95% CI | |
| | | | Lower | Upper | | | Lower | Upper |
| Spontaneous production proficiency (5 points) | | | | | | | | |
| Overall (pre) | 4.2 | .42 | 4.1 | 4.4 | 4.3 | .53 | 4.0 | 4.5 |
| Trained (pre) | 4.0 | .64 | 3.8 | 4.3 | 4.2 | .71 | 3.8 | 4.6 |
| Untrained (pre) | 4.4 | .48 | 4.2 | 4.5 | 4.3 | .47 | 4.1 | 4.6 |
| Overall (post) | 4.3 | .42 | 4.2 | 4.5 | 4.3 | .36 | 4.1 | 4.5 |
| Trained (post) | 4.3 | .53 | 4.1 | 4.5 | 4.2 | .48 | 3.9 | 4.5 |
| Untrained (post) | 4.3 | .65 | 4.1 | 4.5 | 4.5 | .40 | 4.3 | 4.7 |

The normality test (Kolmogorov-Smirnov) showed that there was a significant departure between participants' pre-test scores (trained and untrained) and normal distribution (trained: $D$ = .138, $p$ = .01; untrained: $D$ = .114, $p$ = .026). Data transformation (log) was applied to transform the non-normally distributed data to normality. Pre-test scores were first submitted to a one-way ANOVA with the between-group factor Group. It found no significant Group effect ($p$ > .05), suggesting that there was no pre-existing group difference at the beginning of the project. A three-way ANOVA suggested that the Group × Lexis × Time interaction was significant ($F$ (1, 47) = 4.480, $p$ = .04) with a medium effect size, $\eta_p^2$= .087. Bonferroni multiple comparisons revealed that the experimental group significantly improved their production of the target sounds in trained lexical contexts with moderate effect size ($M$ = 4.0 →4.3, $p$ =.038, $\eta_p^2$ = .089. but not in untrained lexical contexts (p > .05). At the time of pre-test, the difference between rating scores of the trained and untrained items in the experimental group was significant (trained $M$ = *4.0* vs. untrained $M$ = *4. 4*, $p$ =.016, $\eta_p^2$ = .116), indicating that learners were less proficient in the trained items than the untrained counterparts at a spontaneous production level at the time of pre-tests. No statistically significant interactions were found for the control group.

The results suggest that (a) the experimental group substantially improved their nativelikeness of the English [i] and [ɪ] in trained lexical contexts at a spontaneous production level; (b) more considerable improvements were observed in the trained items, which the participants in the experimental group were less proficient at. This suggests that the communicate focus-on-form approach provide greater boosting effects on unfamiliar lexis.

**Figure 11**

*95% Confidence Intervals and Mean Values of the Learners' Spontaneous Production Rating Scores (1: English [i]—5: English [ɪ])*



*Fidelity of Implementation Analysis*

Recasts were provided for the participants in the Experimental group, who received the one-and-half-hour session in small groups of two-to-four students. A total of 98 recasts were directed to 37 out of 40 participants (*M* = 2.6, ranging from 1 to 5 recasts per participant), and 77 recasts were repaired by participants (72.3% repair rate). Due to technical issues, a recording of one hour and 14 minutes was soundless, and one participant from the Experimental group could not be heard during the teaching session. Therefore, the actual number of recasts and repairs might have been more than what was successfully recorded. Compared to the precursor study, Saito (2013), in which the average number of recasts provided for per participant and the repair rate were 17.8 and 91.4% respectively, the repair rate of the current study is considerably low. The gap between the results of the two studies may be the result of the difference in teaching contexts. For instance, the treatment sessions in Saito's (2013) study were carried out face-to-face, whereas the treatment sessions of the current study were implemented via synchronous videoconferencing.

### AP, WM, and L2 Speech Learning

The next objective of the statistical analyses was to further examine the extent to which the improvement of the experimental group could be tied to participants' AP and WM. According to the results of the Kolmogorov-Smirnov test, whereas AP was comparable to normal distribution (*D* = .101, *p* = .780), forward and backward span demonstrated significant deviation, *D* = .276, *p* < .001. Following the analyses of the aptitude-treatment interaction in previous studies (e.g., Chandrasekaran et al., 2010), the aptitude variables (AP and WM) was transformed into categorical variables (i.e., low- and high-aptitude). See Table 8 for summary. Due to multiple comparisons, the alpha value was set to p = .025 via Bonferroni corrections.

**Table 8**

*Summary of AP and WM*

| **Perception** | Experimental (*n* = 39) | | | | Control (*n* = 16) | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 95% CI | | | | 95% CI | |
| Aptitude factors | *M* | *SD* | Lower | Upper | *M* | *SD* | Lower | Upper |
| AP (z scores)[a] | .02 | .68 | -.18 | .22 | -.05 | .56 | -.35 | .24 |
| Forward digit span (10 points) | 9.18 | 1.00 | 8.86 | 9.50 | 9.38 | .62 | 9.05 | 9.70 |
| Backward digit span (9 points) | 8.21 | 1.38 | 7.76 | 8.65 | 8.31 | 1.35 | 7.59 | 9.03 |
| **Production** | Experimental (*n* = 33) | | | | Control (*n* = 16) | | | |
| | | | 95% CI | | | | 95% CI | |
| Aptitude factors | *M* | *SD* | Lower | Upper | *M* | *SD* | Lower | Upper |
| AP (z scores) | .04 | .83 | -.26 | .33 | -.07 | .52 | -.35 | .20 |
| Forward digit span (10 points) | 9.24 | .97 | 8.90 | 9.59 | 9.38 | .62 | 9.05 | 9.70 |
| Backward digit span (9 points) | 8.33 | 1.24 | 7.89 | 8.77 | 8.31 | 1.35 | 7.59 | 9.03 |

*Note*. [a] lower values for more precise AP

## AP and L2 Speech Learning

### *AP vs. Perception*

A total of 39 participants in the experimental group were divided into two subgroups, high-audition (*n* = 20; *M* = -0.43, *SD* = 0.28, *Range* = -0.96 to -0.01) vs. low-audition (*n* = 19; *M* = 0.50, *SD* = 0.49, *Range* = 0.06 to 2.00) by using the group's median values as a cut-off point. See Table 8 for descriptive data. According to the results of multiple comparison analyses, the high- and low-aptitude participants' performance in the perception task was comparable at the time of the pre-tests, *F* (1, 37) = 2.826, *p* = .101, $\eta_p{}^2$ = .071. Yet, the high-audition participants significantly not only enhanced their accuracy scores over time with large effects (*M* = 73.7% → 80.0%, *F* (1, 37) =11.470, *p* = .002, $\eta_p{}^2$ = .237) but also outperformed the low-audition participants at the post-tests (*M* = 80.0% vs. 68.9%, *F* (1, 37) = 5.574, *p* = .024, $\eta_p{}^2$ = .131). For a visual summary, see Figure 12.

### *AP vs. Production*

The participants in the experimental group who successfully recorded their speech (*n* = 33) was divided into high-audition (*n* = 16; *M* = -0.49, *SD* = 0.21, *Range* = -0.84 to -0.19) vs. low-audition (*n* = 17; *M* = 0.52, *SD* = 0.90, *Range* = -0.06 to 3.61) using the median as a cut-off point. The descriptive data was summarized in Table 8. At the time of the pre-test, multiple comparison analyses indicated that the high- and low-audition participants were comparable in terms of producing English [i] and [ɪ] in the controlled task (*M* = 4.3 vs. 4.4, *F* (1, 31) = .013, *p* = .909). The subgroups' performance remained comparable after the instructional treatment (high vs. low *M* = 4.5 vs. 4.5, *F* (1, 31) = .037, *p* = .849). Similarly, both low- and high-audition groups were at the same proficiency level of English [i] and [ɪ] in the spontaneous production task at the pre-test (high vs. low *M* = 4.2 vs. 4.2, *F* (1, 31) = .135, *p* = .716)) and the post-test (high vs. low *M* = 4.3 vs. 4.3, *F* (1, 31) = .089, *p* = .767). For a visual summary, see Figure 13 for controlled production and Figure 14 for spontaneous production.

In summary, while low AP abilities hindered learners' learning gains in perception from the intensive short-time meaning-oriented form-focused instruction, individual differences in AP levels did not set their production gains apart.

**Figure 12**

*95% Confidence Intervals and Mean Values of the Learners' Perception Scores as per AP and WM Conditions*

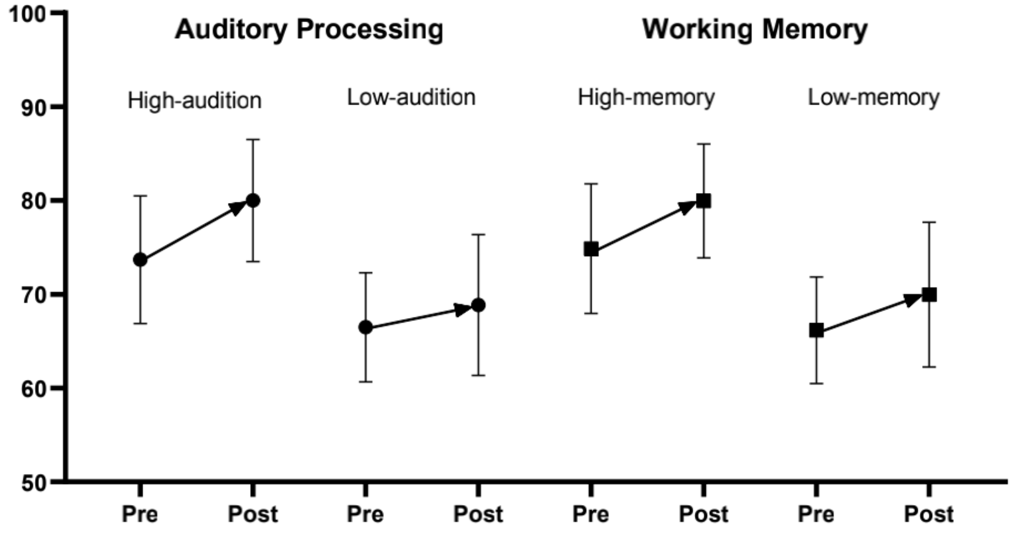


**Figure 13**

*95% Confidence Intervals and Mean Values of the Learners' Controlled Production Scores as per AP and WM Conditions*

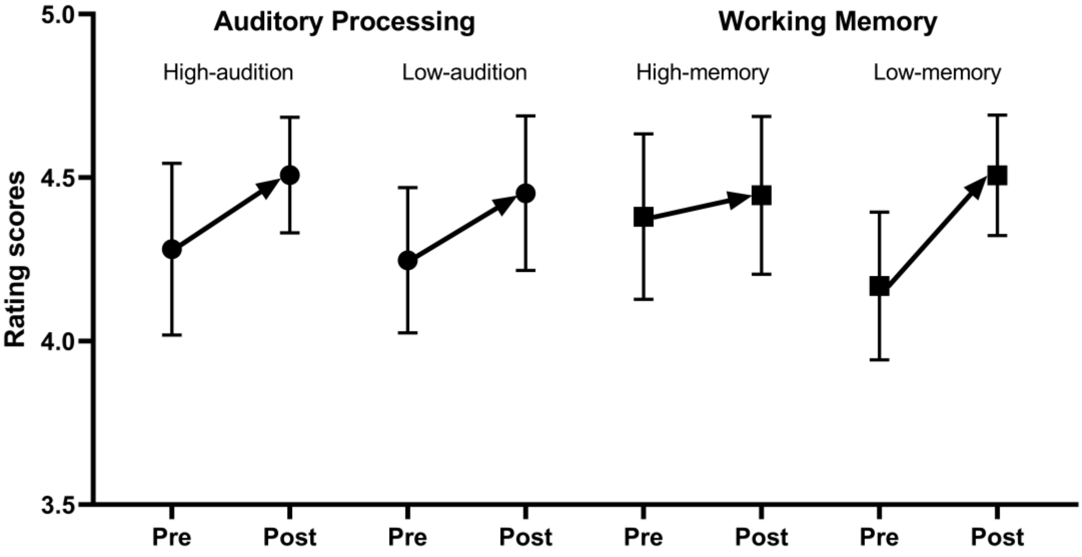
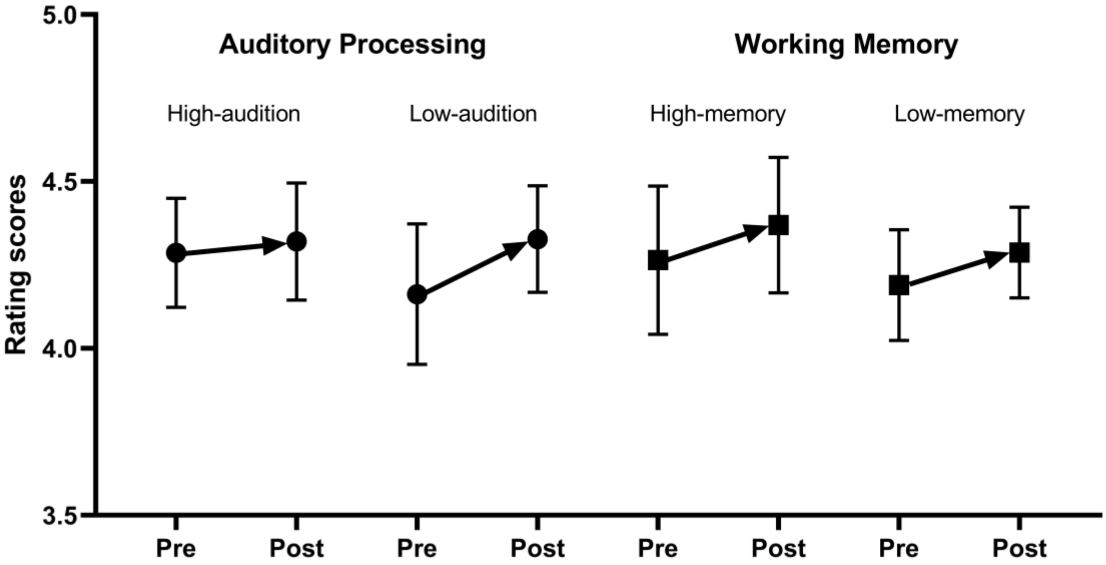


**Figure 14**

*95% Confidence Intervals and Mean Values of the Learners' Spontaneous Production Scores as per AP and WM Conditions*



## Phonological WM and L2 Speech Learning

*WM vs. Perception*

Given that both forward and backward span tasks were significantly correlated, a decision was made to use both scores to subdivide the 39 experimental participants. Those who obtained full scores in forward span (10 out of 10) and backward (9 out of 9) were categorized as high-memory group (*n* = 18). The remaining participants were clustered as low-memory group (*n* = 21). See Table 8 for summary. The results of multiple comparison analyses demonstrated that the high-memory group's vowel performance was marginally higher than the low-memory group at both pre-tests (*M* = 74.9% vs. 66.2%, *F* (1, 37) = 4.247, *p* = .046, $\eta_p^2$ = .103) and post-tests (*M* = 80.0% vs. 70.0%, *F* (1, 37) = 4.343, *p* = .044, $\eta_p^2$ = .105) regarding perception. The participants' learning gains were significant among the high-memory group with medium effects (*M* = 74.9% → 80.0%, *F* (1, 37) =6.391, *p* = .016, $\eta_p^2$ = .147) but marginal among the low-memory group with small effects (*M* = 66.2% → 70.0%, *F* (1, 37) =4.148, *p* = .049, $\eta_p^2$ = .101). For a visual summary, see Figure 12.

*WM vs. Production*

In the same manner of perception analyses, participants who obtained full scores in forward span and backward were categorized as high-memory group (*n* = 15) with the remaining participants categorized as low-memory group (*n* = 18). Table 8 demonstrates the summary of the descriptive data.

The results of multiple comparison analyses demonstrated that the low-memory group's pre-test scores for the controlled production was marginally lower than the high-memory group (*M* = 4.1 vs. 4.5, *F* (1, 31) = 4.830, *p* = .036, $\eta_p^2$ = .135). After receiving the one-and-half-hour communicative form-focused intervention, the low-memory subgroup made marginally larger improvement than the high-memory subgroup (*M* = 0.4 vs. 0, *F* (1, 31) = 4.020, *p* = .054, $\eta_p^2$ = .002). At the time of post-test, the low- and high-memory groups' controlled production proficiency was at the same level, *M* = 4.5 vs. 4.5, *F* (1, 31) = .067, *p* = .797, $\eta_p^2$ = .002.

In the spontaneous production condition, the low- and high-memory subgroups did not differ from each other significantly at in their proficiency of English [i] and [ɪ] (Mean score high vs. low = 4.3 vs. 4.1, *p* = .370, $\eta_p^2$ = .026) before the intervention, and their performance almost stayed same after receiving the 90-minute form-focus instruction (Mean score high vs. low = 4.3 vs. 4.2, *p* = .320, $\eta_p^2$ = .023). For a visual summary, see Figure 13 for controlled production and Figure 14 for spontaneous production. In summary, phonological WM level did not predict the learning gains of the learners' production of English [i] and [ɪ] at controlled and spontaneous levels.

**Relationships between AP and Phonological WM**

The present study also aims to investigate the relationship between AP abilities and phonological WM capabilities. For this purpose, Pearson's correlation was conducted to the whole sample. The overall AP threshold and the overall WM scores did not correlate with each other (*p* > .05). The output in Table 9 shows that, while the measures for each language aptitude were internally correlated with moderate to strong correlations (AP spectral and AP temporal *r* = .339, *p* = .011; DSF and DSB *r* = .402, *p* = .002), there was no significant correlations between the AP measures and the phonological memory measures (see also Table 5 for descriptive statistics). These results indicate that learners with higher spectral processing abilities also had higher temporal processing abilities, and the same pattern applied to the size of PSTM store and the phonological WM abilities. More importantly, the results support the assumption that the AP and phonological memory measures used in the current study measured two different kinds of language aptitudes that did not overlap with one another and function separately in L2 learning. This suggests that the AXB psychoacoustic discrimination tasks and the visual text-entry DS tasks are effective in avoiding tapping into the same perceptual-cognitive mechanism when examining FL aptitude related to L2 speech development.

**Table 9**

*Correlations between Each AP and Phonological WM Measures*

| | AP Spectral | AP Temporal | DSF | DSB |
|---|---|---|---|---|
| AP Spectral | 1 | .339* | .022 | -.008 |
| AP Temporal | .339* | 1 | .013 | -.071 |
| DSF | .022 | .013 | 1 | .402** |
| DSB | -.008 | -.071 | .402** | 1 |

*Note.* *Correlation is significant at the 0.05 level (2-tailed). ** Correlation is significant at the 0.01 level (2-tailed).

## Discussion

### FFI and Recasts and L2 Pronunciation Learning

The first research question asked how could communicative FFI help Chinese learners of English improve their L2 vowel acquisition (English [i] and [ɪ]). Statistical analyses revealed that the 1.5 hours of instruction significantly helped improve L2 vowel perception with medium-to-large effects in both trained and untrained lexical items— learners in the experimental group improved by 4.6% overall in trained and untrained lexical contexts one day (cf. Saito, 2013, for an average increase of 7—10% after a four one-hour training; Lee & Lyster, 2016, for an average increase of 13.7 % after five one-hour training). The 90-minute communicative form-focused session also led to substantial improvements at a spontaneous production level, which, however, was limited to the trained items (mean gains = 0.3 on a 5-point scale). Interestingly, both the experimental and controlled groups' improvement (experimental vs. control = 0.8 vs. 0.4 on 5-point scale) in the untrained lexical items reached statistical significance. The experimental group improved substantially in the absolute value of the scores, but there was no significant difference with the control group because the control group also improved greatly, although only half as much as the experimental group. The learning gains of the control group could be due to the practice effects or test-retest effects— the gains of test performance are resulted from taking the same tests or an alternative test with the same difficulty level multiple times (Roediger & Butler, 2010). In this connection, the treatment session and test-retest effects might have jointly contributed to improved performance of the experimental group in the controlled production tasks.

There are noticeable variations in the effectiveness of the form-focused treatment in three proficiency dimensions—the participants' learning gains was considerable in both trained and untrained lexical conditions in perception but limited to the untrained condition in controlled production and the trained condition in spontaneous production. This might be explained by the *Information Processing Theory by* Miller (1956), where L2 learning is viewed as skill learning (McLaughlin, 1987). There are three phases of L2 acquisition under the information processing framework: declarative knowledge (i.e., input), restructuring (where learners internalize and gain increasing control the knowledge), and automaticity (where the skill becomes automatic) (McLaughlin, 1987). See Table 10 for detailed information.

Learners in the experimental group were most likely in the process of restructuring to automaticity at the time of post-test. First, the significant pre-to-post-test improvement regardless of lexical conditions in the perception tests signifies a completion or near-completion of the proceduralization or creation of perceptual representations of English [i] and [ɪ]. Second, the substantial learning gains observed at a controlled production level is in line with the restructuring or the process between restructuring and automaticity, also known as "controlled processing" (McLaughlin, 1987, p. 135), which is relatively easy. As the main instructional effects was found in the untrained lexical context, it is likely that learners in the experimental group were able to apply their knowledge in a novel situation, indicating their ability of generalization. The spontaneous production requires automaticity, or automatic processing, which is a "learned response…over many trials" (McLaughlin, 1987, p. 134) and occurs after the earlier controlled processing. It is hypothesized that the production of a language-specific speech sound is a reflection of the phonetic category established in LTM (Flege, 1995). The short instructional time and the short interval between the treatment and the post-test may not be sufficient for participants to practice repeatedly or to discern the phonemic differences for the form new phonetic categories to be established in LTM.

**Table 10**

*A Synthesis of Miller's (1956) IPT and Skehan's (2016) Micro Aptitude Models*

| Developmental Phrases | L2 acquisition processes | Aptitude constructs |
|---|---|---|
| Declarative knowledge | Input processing<br>Noticing<br>Handling feedback [a] | Attentional control<br>Working memory<br>Phonetic coding ability |
| Restructuring | Pattern identification<br>Complexification<br>Handling feedback [a]<br>Error avoidance | Working memory<br>Language analysis ability |
| Automaticity | Automatization<br>Creating new repertoire, achieving salience<br>Lexicalizing | Retrieval memory<br>Chunking[b] |

*Note.*

[a] 'Handling feedback' was classified as 'restructuring' by Skehan (2016). Since recast is an input-providing corrective feedback (Sheen, 2011), it is also categorized as a kind of input.

[b] 'Chunking' is the ability to bridge STM and LTM (Ellis, 1996).

One important implication of this part of the findings is that when teaching L2 sounds that are phonetically similar to learners' L1 sounds to adult learners or helping them distinguish similar L2 sounds, recasts and FFI with explicit instruction is an effective remedy. However, in order to achieve long-term enhancement, such remedy would need to be consistently provided in language classrooms for an extended period of time.

### AP, Phonological Memory and L2 Speech Learning

In comparison to learners in the low-audition subgroup, learners with relatively high AP abilities made significantly greater progress on the perception of English [i] and [ɪ]. This finding broadly supports the findings of previous studies examining the association between domain-general AP and L2 pronunciation development (e.g., Saito et al., 2020a; Saito et al., 2020b; Zheng et al., 2020): domain-general AP is a significant predictor of L2 pronunciation acquisition in a naturalistic/L2 immersion context. The aptitude-instruction interaction found between audition levels and communicative focus on form indicates that the significant effects of domain-general AP on L2 pronunciation also exist in instructed learning conditions, which had not been reported before (Saito et al., 2021). Participants' audition levels did not predict their learning gains in production tasks. For the majority of people, the auditory channel is the first channel through which they receive language input (Kachlicka et al., 2019). It can therefore be assumed that AP is the 'gateway-like' aptitude that determines how the oral instruction (e.g., recasts) on L2 sounds are perceived before they can be stored in memory or prepared for production, and that the AP mechanisms may not be the devices responsible for production. This finding shed some light on Skehan's (2016, see Table 10) micro prospective of aptitude, where specific aptitude constructs are related to specific processing stages of L2 acquisition. According to the findings here, it is proposed that domain-general AP is responsible for *input processing* at *declarative knowledge* stage.

Learners' phonological WM only exerted limited mediating effects on the effectiveness of communicative form-focused instruction in the current study. The high-memory subgroup made larger gains in perception than the low-memory subgroup with marginal statistical significance, suggesting that while phonological WM may predict L2 learners' perception gains from FFI and recasts. This is in line with the recent finding of Lee (2021) that WM was a significant predictor of learning success when they received proactive focus-on-form activities on phonological aspects of L2 morphology. Surprisingly, the low-memory subgroup demonstrated larger gains than the high-memory subgroup in the controlled production tests over time. This contradicts the argument that smaller phonological WM capacity leads to increased L1 interference (Darcy et al., 2015). Although counterintuitive, the finding in question might be partly explained by the characteristics of the FFI + recast treatment. For learners with relatively smaller phonological WM capacity, who tend to be less sensitive to phonological information (Knoop-van Campen et al., 2018), FFI might have helped to increase their phonological awareness of words with the target sounds, and recasts might have further helped them to notice the target sounds, thus establishing new phonetic categories at a segmental level (Saito, 2013). This could also be related to the ceiling effect of the high-memory subgroup. As mentioned in the previous section, the high-memory subgroup showed a stagnant pre- to post-test trend in spontaneous production test scores (*M* = 4.3→ 4.3). It is possible that, for these L2 learners who had never had immersion experience in an English-speaking country longer than a month, they had already reached the highest accuracy level possible of pronouncing English [i] and [ɪ] in a spontaneous production condition before the experiment began.

Extending the existing evidence on the ATI effects, the current study further added that two different types of domain-general abilities (AP and verbal WM) mediate the effectiveness of more meaning-oriented, communicatively-authentic instruction. Quoting Skehan (2015), “higher aptitude seems to help a focus-on-form to be ‘dug out’” (p. 14). By contrast, low aptitude may seriously hinder processing, noticing and generalizing information embedded in communicative FFI. Therefore, for leaners with low aptitude, particularly low AP and phonological WM abilities, remedial interventions (e.g., provision of explicit metalinguistic instruction and feedback) may be able to guide them to notice, understand and internalize linguistic targets (Skehan, 2016). As such, learners are not required to ‘dig out’ the ‘hidden’ information from instruction, which demands much less on learners’ perceptual and cognitive abilities.

### The relationships between AP and phonological memory

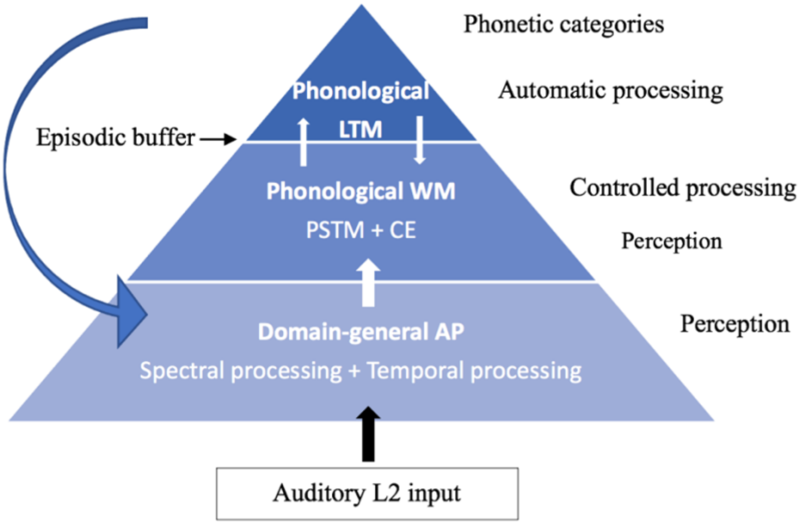
On the question of the relationship between AP and phonological WM, this study found strong intro-aptitude correlations but not inter-aptitude correlations. The result for phonological WM measures lends support for the unity and diversity model by Miyake and Friedman (2012). This model proposes an interdependent relationship between temporary store and executive function. More importantly, this strongly supports the hypothesis that the two aptitude constructs in question are distinctive from each other. The originality of this study is that it is the first time that these two aptitudes, which are critical for L2 speech development, have been investigated together. Thus, in order to differentiate between AP and phonological WM, instead of using the classic sound-based digit span tests, this study utilized the visual-text version. The results demonstrated that using different modalities can effectively avoid different tests tapping into the same area of the perception and cognition. Moreover, interpreted together with the findings for the second research question, AP is more of a first-tier aptitude that determines the quality of L2 sound perception, while phonological WM is a second-tier construct that assists perception, temporarily stores the perceived phonological information and rehearses it for production, controlled production in particular.

The relationships of AP and phonological memory (including LTM) in L2 pronunciation learning is suggested on the basis of the findings of the current study and theory in the existing literature (see Figure 15). The proposed model is modified as per Atkinson and

Shiffrin's (1986) influential model of STM, also known as *modal model*, and Baddeley's (2003) revised multi-component model of WM. As shown in Figure 15, it is assumed that auditory L2 input initially goes through AP mechanisms (i.e., spectral and temporal), where the input is processed according to the established phonetic categories in the phonological LTM. If there were no existing reference, the perception would depend on the learner's AP ability, namely how accurately can the features of the sound be perceived. The preliminarily processed auditory information is then transmitted to the phonological WM system that consists of PSTM and an attentional control construct (i.e., CE) to be rehearsed for further use such as output or for the development of automaticity and new phonetic categories. The interaction between phonological WM and LTM is two-way with the episodic buffer as the interface: the maintenance of WM is predictive of successful formation of LTM (Ranganath et al., 2005), and greater WM facilitates retrieving information from LTM (Cantor & Engle, 1993).

**Figure 15**

*The Processes of L2 Pronunciation Learning and the Involvement of AP and Phonological Memory*



## Conclusion, Limitations and Future Direction

In the context of 55 Chinese speakers' English [i] and [ɪ] acquisition, the current study examined how provision of communicative FFI can facilitate L2 speech learning, and how such instructional gains can be tied to two different types of aptitude factors—i.e., AP and phonological WM. The statistical analyses provided the following primary findings. First, FFI significantly helped improve L2 vowel perception with medium-to-large effects for both trained and untrained lexical items, and spontaneous L2 vowel production with medium effects in for trained items. Secondly, the instructional gains in perception were observed especially among those with a high level of AP (with large effects). Third, learners with high level of phonological WM benefited more from FFI and recasts in the perception and controlled production of L2 vowel sounds than low-memory-level participants, but there were only marginally significant effects. Furthermore, AP and phonological WM were not significantly correlated with each other, suggesting that they are two different types of L2 speech aptitude. The findings echo the previous literature that communicative focus-on-form can significantly facilitate L2 speech acquisition (e.g., Saito, 2013) and that the effectiveness of instruction can be tied to participants' aptitude factors (e.g., Kissling, 2013). An important pedagogical implication is that for L2 learners who are insensitive to new sounds and/or who find memory tasks challenging may need explicit guidance (e.g., metalinguistic explanation) to help them notice and generalize linguistic information embedded in implicit communicative FFI.

This study enriched the research on AP and L2 speech acquisition by extending the line of work to an instructed EFL context. A number of studies have jointly established an essential role of AP in post-pubertal L2 pronunciation learning in naturalistic contexts (e.g., Saito et al., 2020a; Saito et al., 2020b), and the findings of the current study confirmed that it was also true in a stimulated EFL classroom. This study also took the first step to investigate how AP and phonological WM uniquely influence the effectiveness of FFI in L2 pronunciation learning. While AP has been considered as an anchor of L2 speech development in immersion settings, this study provides empirical evidence that both AP and phonological WM play fundamental roles in L2 speech perception in classroom settings. Phonological WM may not be as a strong predictor of instructional gains in the initial stage of L2 speech development (i.e., perception), but it seems to be a cross-domain aptitude factor, involving not only in perception but also the next-level phase (i.e., controlled production). This aligns well the proposed roles of WM in Skehan's (2016) micro aptitude model, but it also led to reconsideration of the model. Given the results of this study and the analysese of the theories, this paper proposes to add AP to the declarative knowledge phase of Skehan's micro aptitude model. To test the hypotheses about AP, WM and L2 speech acquisition, further research that replicates the current study need to be undertaken.

This study could also lead to a number of future directions with a view of achieving a full-fledged picture of the mechanisms underlying successful instructed L2 speech learning.

First, the instruction targeted Chinese speakers' acquisition of English [i] and [ɪ], which could be rated "medium" in terms of learning difficulty because their L1 phonetic inventories have at least one counterpart sound (Chinse [i]). However, existing aptitude literature has shown that aptitude matters especially when it comes to relatively difficult instances of L2 learning. Future studies could use targets that does not have corresponding L1 sounds (e.g., native English speakers' acquisition of Mandarin retroflex affricates). Second, to verify the assumption of the relationship between AP, phonological WM and LTM, further work should include measures of LTM (e.g., available long-term memory, Linck et al., 2013) in a design that lasts for a longer period of time. As stated earlier, LTM is claimed to be responsible for spontaneous produce of acquired L2 knowledge and the establishment of new phonetic categories. Thus, examining correlations between LTM and instructional gains could help develop a more comprehensive understanding of the process of L2 speech learning and the aptitudinal factors involved in each process.

Having said all that, the results regarding WM should be interpreted with caution, as the memory data was collected twice, and the valid data used for analyses was collected after the intervention. The constraints of the global pandemic and geographical distance led to the decision of utilizing the online testing platform, Gorilla. As few published studies made use of the same or similar tests on the platform at the time of this study, there was a lack of practical experience of administering such projects on Gorilla. Therefore, potential improper handling of the tests (e.g., using paper and pen to write the numbers in memory tasks) did not come to the attention of the researcher. In order to accomplish the objectives of the study, the researcher decided to use different tests in the same testing framework and to invigilate the tests via video calls. The repeated administration of aptitude tests may have resulted in test-retest effects. As WM were reported changeable via intensive short-term L2 or memory training (Hayashi et al., 2016), the WM data used in the analyses of this study may actually reflect learners' WM performance that had been trained and improved. Future studies using the same phonological memory tests (i.e., visual-text DSF and DSB) on Gorilla can refer to the note on Gorilla at https://app.gorilla.sc/openmaterials/50646 contributed by this study (Dean, 2021), or the video discussing explicit cheating in online WM tests at https://youtu.be/SwbcqjDAklU (Rodd, 2020) to avoid similar situations.

## References

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/gfz97m

Atkins, P., & Baddeley, A. (1998). WM and distributed vocabulary learning. *Applied Psycholinguistics*, *19*(4), 537-552. https://doi.org/10.1017/S0142716400010353

Atkinson, R. C., & Shiffrin, R. M. (1968). A proposed system and its control processes. In K. W. Spence (Ed.), *The psychology of learning and motivation: Advances in research and theory* (pp. 89—195). New York: Academic Press.

Baddeley, A. D. (2000a). The episodic buffer: A new component of WM? *Trends in Cognitive Sciences*, *4*, 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D. (2000b). Short-term and WM. In E. Tulving &amp; F. I. M. Craik (Eds.), *The Oxford Handbook of Memory* (pp. 77–92). Oxford University Press.

Baddeley, A. (2003). WM: looking back and looking forward. Nature Reviews Neuroscience, 4, 829–839. https://doi.org/10.1038/nrn1201

Baddeley, A. D., & Hitch, G. J. (1974). WM. In G. A. Bower (Ed.), *Recent advances in learning and motivation* (pp. 47–90). New York: Academic Press.

Baills, F., Zhang, Y., Cheng, Y., Bu, Y. & Prieto, P. (2021). Listening to songs and singing benefitted initial stages of second Language pronunciation but not recall of word meaning. *Language Learning*. Advance online publication. https://doi.org/10.1111/lang.12442

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issue in cross-language research* (pp. 171–204). Timonium, MD: York Press.

Cambridge University Press & Assessment (n.d.). *How are language levels described?* https://www.cambridgeenglish.org/learning-english/parents-and-children/information-for-parents/tips-and-advice/011-the-cefr/

Cantor, J., & Engle, R. W. (1993). Working-memory capacity as long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1101–1114. https://doi.org/10.1037/0278-7393.19.5.1101

Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test*. Psychological Corporation.

Chandrasekaran, B., Sampath, P., & Wong, P. (2010). Individual variability in cue-weighting and lexical tone learning. The Journal of The Acoustical Society of America, 128(1), 456-465. https://doi.org/10.1121/1.3445785

Cobb, T. (2012). *The Compleat Lexical Tutor*. http://www.lextutor.ca/vp/

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Routledge.

Darcy, I., Park, H., & Yang, C.L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences, 40*, 63–72. http://dx.doi.org/10.1016/j.lindif.2015.04.005

Dean, P. (2020, March 24). *Digit Span Task (Visual & Text Entry)*. Gorilla. https://app.gorilla.sc/task/4698263

Dean, O. (2021). *Collection of useful tasks and questionnaires*. https://app.gorilla.sc/openmaterials/50646

Derwing, T., & Munro, M. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*, 379–397. https://doi.org/10.2307/3588486

Ellis, N. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*(1), 91-126. https://doi.org 10.1017/S0272263100014698

Ellis, N. C., & Sinclair, S. G. (1996). WM in the acquisition of vocabulary and syntax: Putting language in good order. *Quarterly Journal of Experimental Psychology*, *49*, 234–250. https://doi.org/10.1080/713755604

Elliott, A. (1997). On the teaching and acquisition of pronunciation within a communicative approach. *Hispania*, *80*, 95–108. https://doi.org/10.2307/345983

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, *9*(2), 147–171. https://doi.org/10.1191/1362168805lr161oa

French, L. M., & O'Brien, I. (2008). Phonological memory and children's second language grammar learning. *Applied Psycholinguistics*, *29*(3), 463-487. http://dx.doi.org.libproxy.ucl.ac.uk/10.1017/S0142716408080211

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), Speech perception and linguistic experience: Issue in cross-language research (pp. 233-277). Timonium, MD: York Press.

Goo, J. (2012). Corrective feedback and WM capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition*, *34*(3), 445–474. https://doi.org/10.1017/S0272263112000149

Gooch, R., Saito, K., &amp; Lyster, R. (2016). Effects of recasts and prompts on L2 pronunciation development: Teaching English /ɹ/ to Korean adult EFL learners. *System*, *60*, 117–127. https://doi.org/10.1016/j.system.2016.06.007

Granena, G. (2012). *Age differences and cognitive aptitudes for implicit and explicit learning in ultimate second language attainment* (Publication No. 3517774) [Doctoral dissertation, University of Maryland]. ProQuest Dissertations and Theses Global.

Granena, G. (2013). Reexamining the robustness of aptitude in second language acquisition. In G. Granena & M. H. Long (Eds.), *Sensitive periods, language aptitude, and ultimate L2 attainment* (pp. 179–204).

Granena, G. & Yilmaz, Y. (2019), Corrective Feedback and the Role of Implicit Sequence-Learning Ability in L2 Online Performance. *Language Learning*, *69*, 127-156. https://doi.org/10.1111/lang.12319

Hayashi, Y., Kobayashi, T., & Toyoshige, T. (2016). Investigating the relative contributions of computerised working memory training and English language teaching to cognitive and foreign language development. Applied Cognitive Psychology, 30, 196–213. https://doi.org/10.1002/acp.3177

Hillenbrand, J. M., & Clark, M. J. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*, 3013-3022. https://doi.org/10.1121/1.1323463

Hwu, F., & Sun, S. (2012). *The aptitude-treatment interaction effects on the learning of grammar rules*. *System*, *40*(4), 505-521. https://doi.org/10.1016/j.system.2012.10.009

Indrarathne, B., & Kormos, J. (2018). The role of WM in processing L2 input: insights from eye-tracking. *Bilingualism: Language and Cognition*, *21*(2), 355-374. https://doi.org/10.1017/S1366728917000098

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10* (2), 135-159. https://doi.org/10.1080/15434303.2013.769545

Javitt, D. C., & Sweet, R. A. (2015). Auditory dysfunction in schizophrenia: Integrating clinical and basic features. *Nature Reviews Neuroscience*, *16*(9), 535–550. https://doi.org/10.1038/nrn4002

Jiang, Li., Zhang, L. J., & May, S.  (2019). Implementing English-medium instruction (EMI) in China: Teachers’ practices and perceptions, and students’ learning motivation and needs, *International Journal of Bilingual Education and Bilingualism*, *22* (2), 107—119. https://doi.org/10.1080/13670050.2016.1231166

Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language*, *192*, 15-24. https://doi.org/10.1016/j.bandl.2019.02.004

Kalashnikova, M., Goswami, U., & Burnham, D. (2019). Delayed development of phonological constancy in toddlers at family risk for dyslexia. Infant Behavior and Development, 57, 101327. https://doi.org/10.1016/j.infbeh.2019.101327

Kaushanskaya, M. & Yoo, J. (2011). Phonological short-term and WM in bilinguals’ native and second language. *Applied Psycholinguistics, 34*(2013), 1005–1037. https://doi.org/10.1017/S0142716412000100

Kissling, E. (2013). Teaching pronunciation: Is explicit phonetics instruction beneficial for FL learners? The Modern Language Journal, 97(3), 720-744. https://doi.org/10.1111/j.1540-4781.2013.12029.x

Knoop-van Campen, C., Segers, E., & Verhoeven, L. (2018). How phonological awareness mediates the relation between WM and word reading efficiency in children with dyslexia. *Dyslexia*, *24*(2), 156-169. https://doi.org/10.1002/dys.1583

Kourtali, N.-E., & Révész, A. (2019), The roles of recasts, task complexity, and aptitude in child second language development. *Language Learning*, *70,* 179-218. https://doi.org/10.1111/lang.12374

Lee, A. H. & Lyster, R. (2016), Effects of different types of corrective feedback on receptive skills in a second language: A speech perception training study. *Language Learning*, *66*, 809-833. https://doi.org/10.1111/lang.12167

Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. Language Learning, 60, 309-365. https://doi.org/10.1111/j.1467-9922.2010.00561.x

Linck, J.A., Hughes, M.M., Campbell, S.G., Silbert, N.H., Tare, M., Jackson, S.R., Smith, B.K., Bunting, M.F. & Doughty, C.J. (2013), Hi-LAB: A new measure of aptitude for high-level language proficiency. *Language Learning*, 63, 530-566. https://doi.org/10.1111/lang.12011

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake. *Studies in Second Language Acquisition*, *19,* 37–66. https://doi.org/10.1017/S0272263197001034

Majidi, A.E., Janssen, D. & de Graaff, R. (2021). The effects of in-class debates on argumentation skills in second language education. *System*, *101*, 1—15. https://doi.org/10.1016/j.system.2021.102576

Masoura, V. M. & Gathercole, S. E. (2005). Phonological short-term memory skills and new word learning in young Greek children. *Memory*, *13*, 422–429.

McKinney, B., Ding, Y., Lewis, D. A., & Sweet, R. A. (2017). DNA methylation as a putative mechanism for reduced dendritic spine density in the superior temporal gyrus of subjects with schizophrenia. *Translational Psychiatry*, *7*(2). https://doi.org/10.1038/tp.2016.297

McLaughlin, B. 1987. *Theories of Second-Language Learning*. London: Edward Arnold.

Meara, P. (2005). *LLAMA language aptitude tests*. Swansea, England: Lognostics.

Meisel, J. M. (2011). *First and second language acquisition: Parallels and differences*. Cambridge: Cambridge University Press.

Miller, G. (1956). The magical number Seven, plus or minus two. *Psychological Review*, *101* (2), 343-352.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. A common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, and textbooks across Europe. In I. Bartning, M. Martin & I. Uedder (Eds.), *Communicative proficiency and linguistic development: intersections between SLA and language testing research* (pp. 211—232). European Second Language Association.

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: four general conclusions. *Current Directions in Psychological Science*, *21*(1), 8–14. https://doi.org/10.1177/0963721411429458

Modern language aptitude test (MLAT). (2012). http://lltf.net/aptitude-tests/language-aptitude-tests/modern-language-aptitude-test-2/

Munoz, C. (2014). Contrasting effects of starting age and input on the oral performance of foreign language learners. Applied Linguistics, 35(4), 463-482. https://doi.org/10.1093/applin/amu024

Norris, J.M., & Ortega, L. (2000), Effectiveness of L2 Instruction: A Research Synthesis and Quantitative Meta-analysis. *Language Learning*, *50*, 417-528. https://doi.org/10.1111/0023-8333.00136

O'Brien, I., Segalowitz, N., Freed, B., & Collentine, J. (2007). Phonological memory predicts second language oral fluency gains in adults. *Studies in Second Language Acquisition*, *29*, 557–582. https://doi.org/10.1017/S027226310707043X

Parlak, Ö., & Ziegler, N. (2016). The impact of recasts on the development of primary stress in a synchronous computer-mediated environment. Studies in Second Language Acquisition, 39(2), 257-285. https://doi.org/10.1017/S0272263116000310

Ranganath, C., Heller, A., Cohen, M.X., Brozinsky, C.J. & Rissman, J. (2005). Functional connectivity with the hippocampus during successful memory formation. *Hippocampus, 15*, 997-1005. https://doi.org/10.1002/hipo.20141

Rassaei, E. (2015). Oral corrective feedback, foreign language anxiety and L2 development. System, 49, 98–109. https://doi.org/10.1016/j.system.2015.01.002

Révész, A. (2012). WM and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, *62*(1), 93–132. https://doi.org/10.1111/j.1467-9922.2011.00690.x

Robb, J. [Jennifer R]. (2020, July 2nd). *BeOnline2020: Jenni Rodd - Ensuring data quality when you can't see your participants* [Video]. YouTube. https://youtu.be/SwbcqjDAklU

Robinson, P. (2002). Learning conditions, aptitude complexes and SLA: A framework for research and pedagogy. In P. Robinson (Ed.), *Individual differences and instructed language learning*, 113–133.

Roediger, H., & Butler, A. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

Saito, K. (2013). The acquisitional value of recasts in instructed second language speech learning: Teaching the perception and production of English /ɹ/ to adult Japanese learners. *Language Learning*, *63*, 499-529. https://doi.org/10.1111/lang.12015

Saito, K. (2021). What characterizes comprehensible and native-like pronunciation among English-as-a-second-language speakers? Meta-Analyses of phonological, rater, and instructional factors. *TESOL Quarterly*. https://doi.org/10.1002/tesq.3027

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning, 62*(2), 595-633. http://dx.doi.org/10.1111/j.1467-9922.2011.00639.x.

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. Language Learning, 69(3), 652-708. https://doi.org/10.1111/lang.12345

Saito, K., Sun, H., Kachlicka, M., Alayo, J., Nakata, T., & Tierney, A. (2020b). Domain-general auditory processing explains multiple dimensions of L2 acquisition in adulthood. *Studies in Second Language Acquisition*, 1-30. https://doi.org/10.1017/S0272263120000467

Saito, K., Sun, H., & Tierney, A. (2020a). Domain-general auditory processing determines success in second language pronunciation learning in adulthood: A longitudinal study. *Applied Psycholinguistics*, *41*(5), 1083-1112. https://doi.org/10.1017/S0142716420000491

Saito, K., Suzukida, Y., Tran, M. and Tierney, A. (2021). Domain-general auditory processing partially explains second language speech learning in classroom settings: A review and generalization study. *Language Learning*, *71*, 669-715. https://doi.org/10.1111/lang.12447

Saito, K., Trofimovich, P. & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization study. *Applied Linguistics*, *38*(4), 439–462. https://doi.org/10.1093/applin/amv047

Sheen, Y. (2011). *Corrective feedback, individual differences and second language learning*. New York: Springer.

Skehan, P. (2002). Theorising and updating aptitude. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 69–93). John Benjamins.

Skehan, P. (2015). Foreign language aptitude and its relationship with grammar: A critical overview. *Applied Linguistics*, *36*(3), 367–384. https://doi.org/10.1093/applin/amu072

Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. In G. Granena, D. Jackson & Y. Yilmaz (eds.), *Cognitive individual differences in L2 processing and acquisition* (pp. 15–38). Amsterdam: John Benjamins.

Smith, B. L., Johnson, E., & Hayes-Harb, R. (2019). ESL learners' intra-speaker variability in producing American English tense and lax vowels. *Second Language Pronunciation*, *5*(1), 139–164. https://doi.org/10.1075/jslp.15050.smi

Snowling, M. J., Gooch, D., McArthur, G., & Hulme, C. (2018). Language skills, but not frequency discrimination, predict reading skills in children at risk of dyslexia. Psychological Science, 29(8), 1270–1282. https://doi.org/10.1177/0956797618763090

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. Language Learning, 60(2), 263-308. https://doi.org/10.1111/j.1467-9922.2010.00562.x

Sun, H., Saito, K., & Tierney, A. (2021). A longitudinal investigation of explicit and implicit auditory processing in L2 segmental and suprasegmental acquisition. *Studies in Second Language Acquisition*, *43*(3), 551-573. https://doi.org/10.1017/s0272263120000649

Wang, C. (1988). *The production and Perception of English vowels by native speakers of Mandarin*. Unpublished M.A. thesis, University of Alabama at Birmingham.

Wang, X. (1997). *The Acquisition of English Vowels by Mandarin ESL Learners: A Study of Production and Perception*. Unpublished MA Thesis, Simon Fraser University.

Wang, X. (2002). *Training Mandarin and Cantonese speakers to identify English vowel contrasts: Long-term retention and effects on production.* Unpublished PhD Thesis, Simon Fraser University.

Wang, X. (2006). Mandarin listeners' perception of English vowels: Problems and strategies. *Canadian Acoustics*, *34*(4), 15-26.

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. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences: San Francisco 1999* (pp.125-128). Berkeley, CA: University of California Berkeley.

Wechsler, D. (2008). *Wechsler adult intelligence scale*. Pearson.

Winke, P., Gass, S., & Myford, C. (2013). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231–252. https://doi.org/10.1177/0265532212456968

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of WM capacity and language analytic ability. *Applied Linguistics, 34*(3), 344–368. https://doi.org/10.1093/applin/ams044

Yilmaz, Y., & Granena, G. (2015). The role of cognitive aptitudes for explicit language learning in the relative effects of explicit and implicit feedback. *Bilingualism: Language and Cognition*, *19*(1), 147-161. https://doi.org/10.1017/S136672891400090X

Yilmaz, Y., & Granena, G. (2021). Implicitness and Explicitness in Cognitive Abilities and Corrective Feedback: A double discussion? *Studies in Second Language Acquisition*, *43*(3), 523-550. https://doi.org/10.1017/S0272263120000601

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46. https://doi.org/10.1016/s1364-6613(00)01816-7

Zhang, Y. X., Barry, J. G., Moore, D. R., & Amitay, S. (2012). *A new test of attention in listening (tail) predicts auditory performance*. *PLoS ONE*, *7*(12). https://doi.org/10.1371/journal.pone.0053502

Zheng, C., Saito, K., & Tierney, A. (2020). Successful second language pronunciation learning is linked to domain-general auditory processing rather than music aptitude. *Second Language Research*. https://doi.org/10.1177/0267658320978493

# Appendices

## Appendix A. Teaching Guide

1. Introduction (10 minutes)
    - Self-introduction
    - Main Topic

      Almost all of you have taken the IELTS test. Some have got the scores you wanted or needed, but some are still working on getting better scores. I am frequently asked by my students, “In the speaking test/interview, I often can’t think of anything to say in part 3, what should I do?” As you know, part 3 is double-way discussion, which means you need to discuss some general topics with the examiner. You need to justify your opinion with reasons and examples. Even if you try to get away with just giving your opinion, the examiner will follow up and ask “Why?” “Can you give an example?” Many candidates find it hard to think of anything on the spot. Instead of being caught by surprise, why don’t we foster such a way of thinking and speaking—critical thinking and reasoning? This is argumentative skills, which is the topic of this session. Argumentative skills are not only important in IELTS and academic life, it’s also necessary to survive our daily lives.
    - Contents of the session
        a. How to critique?
        b. How to debate?
        c. Team Debate

2. How to critique? (5 minutes)
    - Explanation

      In order to make good and convincing arguments, we are first going to learn how to find problems in others’ opinions (i.e., critique techniques). It’s very easy. For example, let’s work on a warm-up question.
    - Warm-up Question

**“We should seek help from doctors immediately when we feel sick because it is better for us.”**

a) Introduction & Casual talk

• “-----, could you please read the argument?”

• “Do you seek doctor’s help immediately when you feel sick?” “Do you know anyone who does that?”

b) Asking for students’ opinions (class)

• Is this argument good and convincing? Do you agree with this argument?

• If not, please justify why so?

c) Showing *the example answer*.

• “-----, could you read *Opinion* for us?”

• Corrective feedback is offered when the student mispronounces target words.

3. How to debate? (10 minutes)

- Explanation

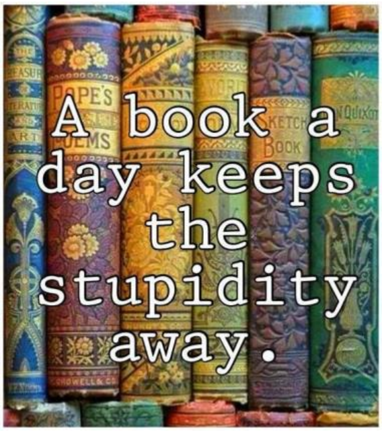
  Once we know how to critique, learning how to debate is even easier! What you need to do is listen to the other party’s speech, find problems in it and carefully and logically critique them. Let’s give it a go.

- Showing critiques of *the example answer*.

  • “-----, could you read *Analogy* for us?”

  • Corrective feedback is offered when the student mispronounces target words.

- Statement 1



**“Reading is the best way to get rid of stupidity.”**

a) Introduction & Casual talk

• “-----, could you read the argument?”

• “Do like reading?” “Do you read a lot?” “What do you usually read?” “Do you think reading makes people smarter?”

b) Asking for students’ opinions (class)

• “Do you agree with this statement?” “Is reading the BEST way to make people less stupid/smarter?”

• If you do think so, please justify why so? If not, are there better ways to get rid of stupidity? Please justify why it is better than reading?

c) Showing the sample answers.

d) Presentation

•Roughly divide the class into 2 opinion groups.

e.g., make ○ gesture if you agree, make ✕ gesture if you disagree. Please make the gesture at the same time when I say “Do you agree or disagree, 123?”

• Ask a student from one team to speak first.

•Ask a student from the opposite team to try to object and justify.

4. Procedures & Rules (5 minutes)

5. Casual debate (15 minutes)

- Statement 2

**"Young people are obligated to give up their seats to the elderly in public transport when the elderly have no place to sit."**

a) Playing videos

- Play the *first* video.
- Ask a student to summarize the video.

  "-----, could you please summarize what happens in the video?"
- Ask the class opinion.

  "What do you think of the old lady's behaviour?"

- Play the *second* video.
- Ask a student to summarize the video.

  "-----, could you please summarize what happens in the video?"
- Ask the class opinion.

  "What do you think of the old gentleman's behaviour?"

b) Introduction & Casual talk

- Young people giving up seats to seniors in public transport vehicles is very commonly seen in China. It is so common that some elders think the young generation is being too kind to them, but some other elders think it's the young people's obligation.
- "-----, could you read the statement for us?"
- "Do you give up your seats in public transport every time you see an old person have no place to sit?" "Have you ever been asked to give up your seat by an old person?" "Do you agree with the statement?"

c) Dividing the class into 2 teams by *"the BLACK & WHITE game"*

d) Brainstorming in breakout rooms (3 minutes)

• Brainstorm with your partner in breakout rooms for 3 minutes, and we'll have a trail debate. IMPORTANT: try to think of at least 2 or 3 ideas and support evidence, each of you has to speak once. Decide who speaks first.

• If you have extra time left, think of some extra ideas, as you don't know what the other party will come up with.

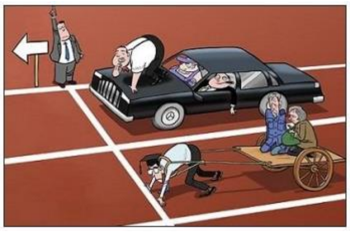
e) Debate (casual)

• Judge & write down points gained by each team.

• Corrective feedback is given when students mispronounce target words.

6. Formal debate (15 minutes)

a) Manners



***"In China, 'impoverished families can hardly nurture rich sons'."***

*Rich: become rich/richer; reach a high/higher social status; fulfil one's dreams*

*Sons: children*

b) Playing the video

c) Summarizing the video (instructor)

Two famous public in China are talking about the phenomenon that the numbers of students from the countryside (where poverty is seen more often) in national key universities are decreasing. They are discussing the reasons behind the phenomenon and also give some advice on what students from the countryside can do to thrive or survive the competition with students from cities.

- Read the statement to the class and explain the key terms (instructor).
- Dividing the class into 2 teams by *"the BLACK & WHITE game"*
- Brainstorming in breakout rooms (5 minutes)

d) Debate (formal)

• Each speech is timed. The time limit is 2 minutes.

• Take notes of each team's point and supporting evidence for judgement.

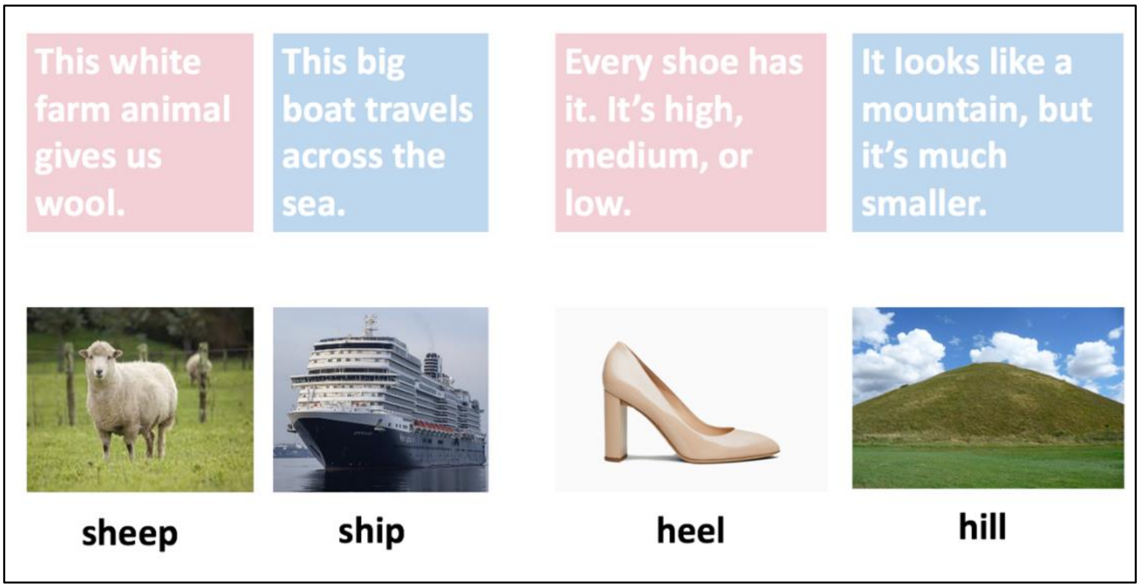
• No corrective feedback is given during the debate.

## Appendix B. Awareness Tasks

### Warm-up Game 1: the riddles (adapted from Kwiatkowska, 2015).

Guess words from the given description.

Kwiatkowska, G. (2015, September 6). *Minimal pairs pronunciation game*. https://www.lessonplansdigger.com/2015/09/06/minimal-pairs-pronunciation-game/



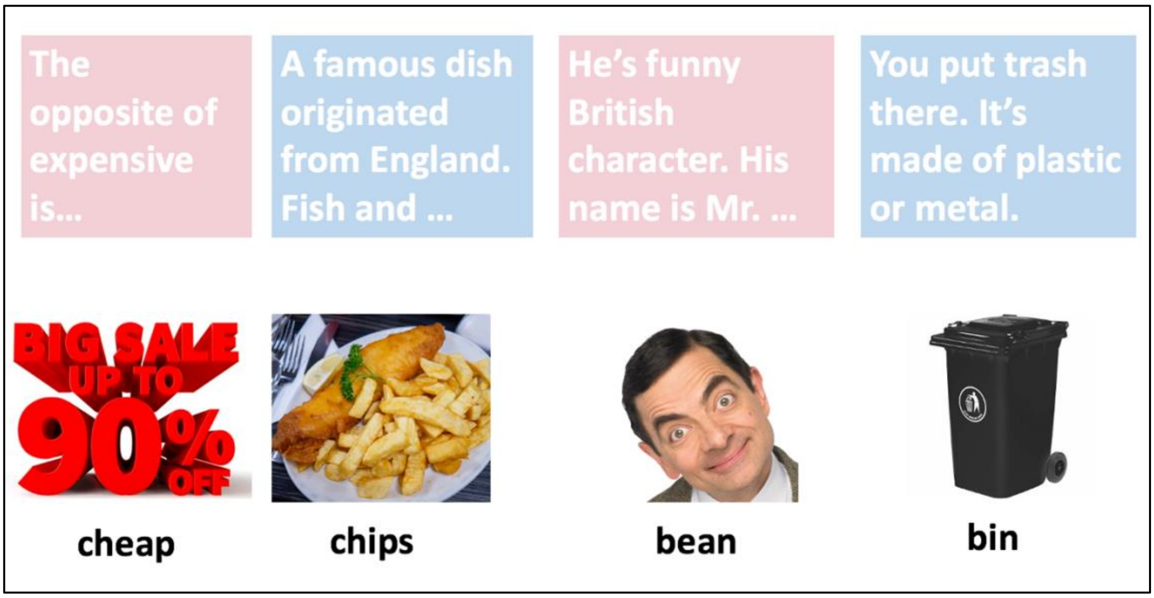
This white farm animal gives us wool.

This big boat travels across the sea.

Every shoe has it. It's high, medium, or low.

It looks like a mountain, but it's much smaller.

sheep ship heel hill



The opposite of expensive is…

A famous dish originated from England. Fish and …

He's funny British character. His name is Mr. …

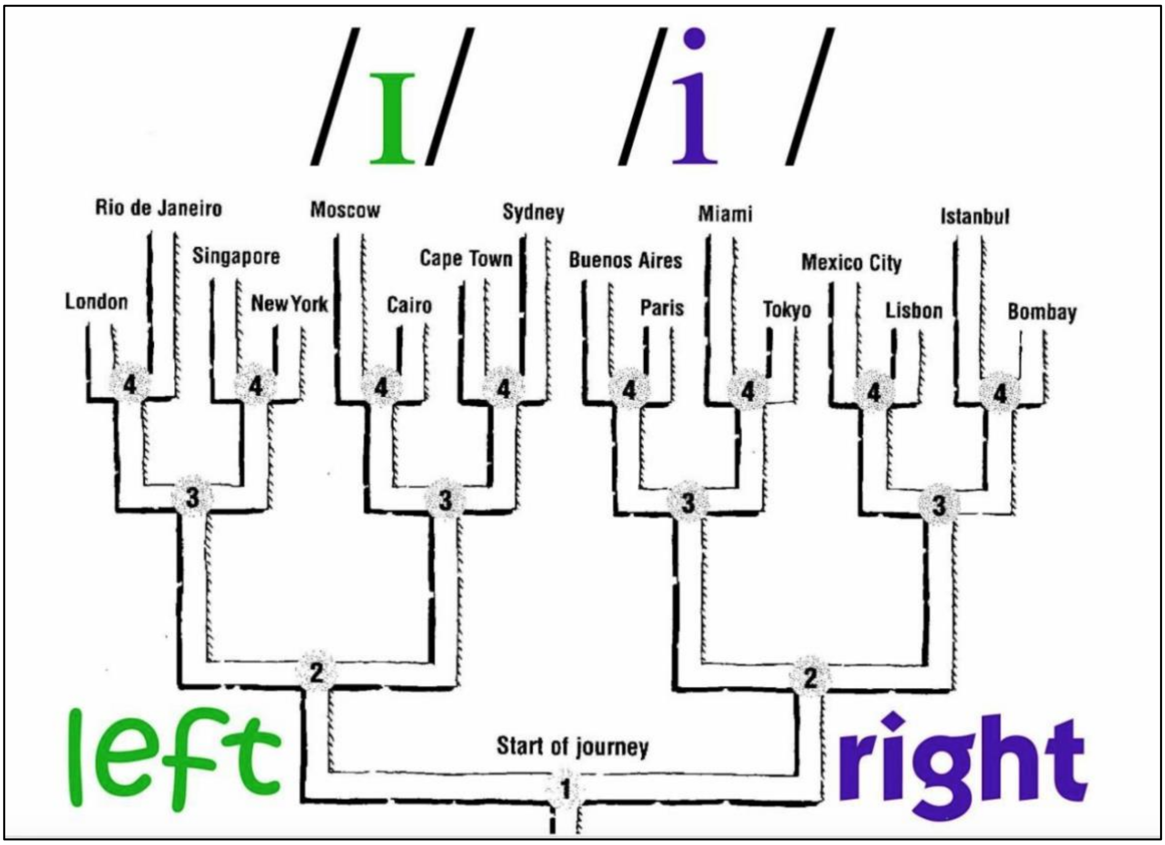
You put trash there. It's made of plastic or metal.

cheap chips bean bin

**Warm-up Game 2: the Pronunciation Journey (Hancock, 2013)**

Take a left when hearing a word with the sound. Take a right when hearing a word with the sound. Check whether your destination matches that of your teacher's.

Handcock, M. (2013, November 19). *Pron Journey hit v heat*. http://hancockmcdonald.com/materials/pron-journey-hit-v-heat



/ɪ/ /i/
Rio de Janeiro
Singapore
London
New York
Moscow
Cairo
Cape Town
Sydney
Buenos Aires
Paris
Miami
Tokyo
Mexico City
Lisbon
Istanbul
Bombay
4
3
2
1
left
Start of journey
right