Shadowing for pronunciation: a systematic review

Benen Nadelek Whitworth

MSc in Applied Linguistics for Language Teaching, 2024

DECLARATION BY THE CANDIDATE AS AUTHOR OF THE DISSERTATION

1. I understand that I am the owner of this dissertation and that the copyright rests with me unless I specifically transfer it to another person.
2. I allow the Department to deposit on my behalf a copy of this dissertation in the Oxford University Research Archive ('ORA') where it shall be freely available online for use in accordance with ORA's Terms and Conditions of Use [https://ora.ox.ac.uk/terms_of_use].
3. I understand that this dissertation should not contain material that can be used to personally identify individuals or specific groups of individuals (unless permission has been obtained from the individuals) and that such material should be removed before this dissertation is deposited in ORA.
4. I agree to be bound by the terms of the ORA Grant of Non-exclusive Licence [https://ora.ox.ac.uk/deposit_agreements] and I warrant that to the best of my knowledge, making my thesis available on the internet will not infringe copyright or any other rights of any other person or party, nor contain defamatory material.
5. I agree that my dissertation shall be available for download in ORA in accordance with paragraphs 2, 3 and 4 above.

Signed:	Benen Nadelek Whitworth
Date:	05/08/2024

Shadowing for pronunciation: a systematic review

Abstract

This dissertation is a systematic review of research into a popular pronunciation teaching technique: shadowing. The review explores the evidence available on the effectiveness of shadowing for improving L2 learners' pronunciation, and learners' evaluations of the technique. Six databases were searched for eligible studies, resulting in a total of 44 studies included after screening. Results from these studies are presented in a narrative synthesis. In general, the evidence presented suggests that shadowing training can help improve learners' comprehensibility, intelligibility, and accentedness, as well as certain aspects of suprasegmental pronunciation control, such as fluency and prosody. Research into the impact of shadowing on segmental pronunciation control was, however, inconclusive. In terms of student perceptions, learners appear to view shadowing as interesting, enjoyable, and effective. However, results are tempered by a number of methodological issues. For instance, in studies exploring pronunciation improvement after shadowing, there was an over reliance on controlled speaking tasks, and a lack of connection between acoustic measurements and real listener judgments. In studies exploring learner perceptions of the technique, there was also a predominance of survey-based research and teacher-collected data. Future studies are advised to address these issues through the inclusion of both spontaneous and controlled tasks, a combination of acoustic measurements and listener judgments, and use of interviews and classroom observation to provide richer qualitative data. In addition, there is a need for research exploring shadowing amongst learners with a wider range of L1s, L2s, and nationalities.

Acknowledgements

Many thanks to my supervisor, Heath Rose, for helping me so attentively at every stage of the process of this dissertation. Thank you also to Filip Bigos, for your assistance with the second screening in this review. And, finally, a huge thanks to my partner, Pablo, and family for all your support and encouragement throughout my studies.

Table of contents

List of tables

List of figures

List of abbreviations

CD - Compact Disc
EFL - English as a Foreign Language
EIKEN - Jitsuyo Eigo Gino Kentei (Test in Practical English Proficiency)
ELF - English as a Lingua Franca
ESL - English as a Second Language
ERIC - Education Resources Information Centre
ESL - English as a Second Language
F0 - Fundamental Frequency
IPA - International Phonetic Alphabet
L1 - First Language
L2 - Second Language
MATLAB - Matrix Laboratory
MSc - Master of Science
PRISMA-P - Preferred Reporting for Systematic Review and Meta-Analysis Protocols
RQ - Research Question
SA - Scaffolded Auditory
US - United States
VA - Visual Auditory
VOT - Voice Onset Time

1. Introduction

1.1. Background to shadowing in pronunciation teaching

Pronunciation is an important component of second language learning and use. Intelligible, comprehensible pronunciation has been shown to be crucial for effective and comfortable communication when speaking a second language (Kelly, 2000/2013). It is therefore an essential part of overall communicative competence in a language (Morley, 1991). In addition, poor pronunciation can lead to a range of problems for learners, such as misunderstandings and negative social judgements (Zoss, 2015), or a decrease in willingness to communicate (Yates, 2001; Zielinski, 2012).

Despite its importance, pronunciation is often neglected in language teacher training programmes (Derwing, 2010) and in teaching practice (Foote et al., 2016). Indeed, research indicates that, when interviewed, both native and non-native speaking language teachers often have many questions about what and how to teach in this area (Couper, 2020).

In such a context, research exploring the effectiveness of different pronunciation training techniques, to guide both language instructors and students, is essential. Shadowing is one technique currently enjoying a boom in popularity (Hamada, 2019a). It involves listening to a short audio text, without a script, and repeating what is heard as simultaneously as possible. Importantly, shadowing involves repeating part of an utterance whilst also listening to incoming parts of the same utterance. This makes it different to traditional listen-and-repeat exercises, which involve a listening period, followed by a pause in which learners repeat what they have heard.

Whilst shadowing is currently popular, and research into its contribution to developing L2 listening skills is robust (Hamada, 2021), research into its effectiveness for training pronunciation has more been limited, and less systematic and conclusive (Foote, 2017; Hamada, 2019a). For this reason, it is a timely moment to consolidate current research in the area via a systematic review.

1.2. Aim of the dissertation

Using systematic review methodology, this dissertation explores the current research landscape surrounding use of the technique of shadowing for pronunciation training. More specifically, it explores the effectiveness of shadowing for improving L2 pronunciation, and learners' assessments of the technique. Its aim is to identify and synthesise the present state of empirical research related to these questions, as well as to assess the quality of included studies and provide recommendations for future research in the field.

1.2. Overview of the dissertation

The dissertation is organised in the following way. Through a narrative review of related literature, Chapter 2 provides a more detailed account of the aforementioned rationale for the present systematic review. It expands upon the importance of pronunciation teaching and the need for research on different pronunciation teaching techniques. It also defines shadowing and explores the historical growth and recent boom in use of the technique, before presenting the research gaps in the shadowing literature that the review aims to fill. Based on this information, two guiding research questions are presented. Chapter 3 presents the methodology used in the study, providing details on the review protocol, eligibility criteria, selection of databases, development of search strings, and selection of studies. It also details the development of the data extraction grid. Chapter 4 presents the results of 44 eligible studies, organised around two themes: pronunciation improvement, and learner perceptions of shadowing. Chapter 5 provides a discussion of these results, identifying key methodological and theoretical limitations of the sample, and how these issues may be addressed in future research in the area.

2. Literature review

This literature review further develops the case for a systematic review of research into shadowing for pronunciation. It aims to situate shadowing within the wider landscape of pronunciation teaching and learning, highlighting it as a pronunciation training exercise worth researching. It first outlines operational definitions of "good" pronunciation and why it is important to learners at a communicative, social, and psychological level. It then highlights that, despite the importance of pronunciation, instruction in this area of language knowledge is frequently overlooked in both teacher training and classroom practice, with many teachers feeling underprepared to address it in their classrooms. Given this situation, it argues that research into the effect of pronunciation techniques, and how students evaluate them, is essential. The review presents shadowing as one such technique, and defines it in contrast to other similar exercises like classical listen-and-repeat. The recent "boom" in use of shadowing, and the current gaps in research on the technique are used to argue that the time has come for a systematic examination of related outcomes in the research literature.

2.1. Pronunciation teaching

2.1.1. The importance of pronunciation in language learning and teaching

Definitions of "good" pronunciation have varied across time periods and according to teaching methods and approaches (Hişmanoğlu, 2006). Currently, influential definitions of "good" pronunciation have shifted away from the Nativeness Principle (Levis, 2005, 2020), which holds that attaining a near-native accent should be the goal or model in pronunciation learning and teaching, with *accentedness* as the most important measure of achievement. Instead, many researchers now advocate for the Intelligibility Principle (Levis, 2005, 2020). This principle emphasises that a near-native accent is not a prerequisite for successful communication, which can and does happen between accented speakers. Furthermore, as developmental and age-related factors make attaining a near-native accent impossible for most adult L2 learners, proponents of the Intelligibility Principle argue that focusing on *accentedness* alone can be setting an unachievable goal for most learners. Because of this, many scholars maintain that the key focus and goal for pronunciation instruction should be increasing *intelligibility* (the degree to which speakers are understood) and *comprehensibility* (the ease with which speakers are understood) (Levis, 2005, 2020). These measures are partially independent from accentedness (Derwing & Munro, 2009)

Pronunciation is important for learners on several levels. Firstly, it has been highlighted as a key component of overall communicative competence in an L2 (Morley, 1991). Indeed, both anecdotal and empirical evidence suggested that if learners do not meet a basic threshold of pronunciation control, they are likely to experience problems when communicating orally (Celce-Murcia et al., 1996/2006, p. 7). This can be the case even if learners have a good control of grammar and lexis (Kelly, 2000/2013; Thomson & Derwing, 2015).

In addition, on a psychological level, good pronunciation control may be linked with a positive self-concept in an L2 (Talebzadeh & Gholami, 2015). Studying pronunciation and seeing improvements in this area may also increase students' motivation (Vitanova & Miller, 2002) and confidence in their language abilities (Pourhosein Gilakjani, 2012; Varasarin, 2007; Zoss, 2016). On a more social level, good pronunciation can contribute to a sense of identification with (Edwards et al., 2021), and respect from, L2 communities and peer groups (Derwing, 2003).

Learners themselves often cite pronunciation as a factor contributing to communication problems (Derwing & Rossiter, 2002; Derwing, 2003), sometimes leading to confusion, embarrassing misunderstandings (Vitanova & Miller, 2002), and social judgement (Kang & Ruben, 2014). In addition, limited pronunciation skills can negatively affect interlocutors' assessments of learners' overall language ability (Pourhosein Gilakjani, 2012). These kinds of social judgement and barriers to communication can have negative impacts on learners' professional and social lives living in L2 communities (Zoss, 2015). Moreover, negative experiences with pronunciation can also have a serious impact on some learners' willingness to communicate (Yates, 2001; Zielinski, 2012), creating a vicious cycle in which they have less language practice, and are less likely to improve.

Given the communicative, social, and psychological importance of pronunciation, and the fact that learners frequently express an interest in improving their control of an L2 sound system (Derwing 2003), including pronunciation work in L2 classrooms is of great importance, Moreover, various reviews suggest that classroom pronunciation instruction is effective. Lee et al. (2015), for example, conducted a meta-analysis of 86 quantitative, quasi-experimental studies involving pronunciation instruction. Overall, aggregated results showed a medium to large effect size for pronunciation interventions (d = 0.89 and 0.80 for N-weighted within- and between-group comparisons, respectively), with larger effects for

longer interventions and interventions involving feedback. Effects across different pronunciation features targeted (e.g., segmental vs. suprasegmental features) were relatively homogenous. This confirmed results from a smaller, more exploratory research synthesis (Saito, 2012). In addition, Thomson and Derwing (2015) conducted a narrative review of 75 studies, which was designed to overlap with Lee et al.'s (2015) meta-analysis in terms of study selection. Their results indicated that 82% of studies reported significant improvements in pronunciation, including both segmental and suprasegmental improvements. Overall, these results suggest that pronunciation instruction is beneficial for learners, although the authors do note that the results are somewhat tempered by a focus on a limited subset of pronunciation features in studies, and some predominance of speaking tasks limited in ecological validity (e.g., read aloud tasks).

2.1.2. Pronunciation in teacher training and practice

Despite the importance of pronunciation for learners, and the improvements that pronunciation instruction can yield, pronunciation teaching is often neglected in language programmes (Derwing, 2010). It has also been a teaching practice difficult to situate within the paradigm of communicative language teaching as a whole (Levis, 2005), due to its focus on individuals sounds and prosody (Foote et al., 2016).

Surveys with teachers in a range of countries indicate that a significant percentage of professionals claim to incorporate pronunciation instruction as a regular part of their teaching practice (Foote et al. 2011; Murphy, 2011; Burgess & Spencer, 2000). However, teachers' self-reports may not capture actual classroom behaviour, as many studies involving classroom observation indicate that pronunciation instruction is often infrequent in standard teaching practice (Darcy et al., 2021; Foote et al., 2016; Shah et al., 2017; Wahid & Suhlong, 2013). In addition, research indicates that the pronunciation instruction that does occur is often reactive, involving ad-hoc, incidental corrective feedback of students' errors, rather than planned instruction (Buss, 2013; Couper, 2016, 2019; Foote et al., 2016; Nguyen et al., 2020).

Some studies suggest that the lack of planned, proactive pronunciation instruction amongst teachers may be partly connected to curricula and assessment frameworks that overlook this aspect of language, and a lack of suitable materials and resources (Macdonald, 2002).

Teachers may also be limited by textbooks they are required to use, which may place greater or lesser importance on pronunciation. Popular ESL textbooks, for example, vary greatly in their coverage of pronunciation features and activities and in the quality and types of tasks they include (Derwing et al., 2012)

A lack of attention to pronunciation in the classroom may also be connected to the training that teachers themselves have received. Research indicates that explicit training in how to teach pronunciation is often absent from or limited in ESL teacher training programmes (Murphy, 2014; Derwing, 2019). In addition, when training programmes do offer components related to phonetics and phonology, they often do not focus on how to practically apply this knowledge to teach pronunciation to learners (Murphy, 2014). This leaves teachers unclear as to how to apply their knowledge base in the classroom (Derwing, 2019). Furthermore, surveys (Breitkreutz et al., 2001; Foote et al., 2011) and interviews (Baker, 2011) with teachers have highlighted not only a lack of training in pronunciation instruction, but also dissatisfaction with the quality of training received (Henderson et al., 2012). Many teachers indicate they would like more training opportunities available (Murphy, 2014), and often pose questions about what techniques and strategies are effective (Couper, 2020).

Overall, research suggests a need for both more reactive and proactive pronunciation instruction, ranging from on-the-spot remedial activities and integrated pronunciation activities in listening and speaking classes (Derwing, 2019), to full stand-alone pronunciation classes (Kelly, 2000/2013). In such a context, research exploring the effectiveness of different pronunciation techniques is of great importance. The next section outlines one such technique: shadowing.

2.2. Shadowing

2.2.1. What is shadowing?

Shadowing is a technique originally developed to train beginner interpreters for the cognitively challenging task of listening and speaking simultaneously (Lambert, 1992). The technique involves listening to an audio text, without a script, and repeating what is heard as simultaneously as possible. The first published accounts of the application of the technique in a language learning context were in the 1990s, when Tamai (1992, 1997) applied the method to help train listening skills amongst Japanese learners of English. Since then, interest in shadowing has increased in L2 instruction, with the technique now used in classrooms in a wide range of countries around the world (Hamada, 2019a). In addition, many language learning websites and blogs now recommend shadowing as a way to improve pronunciation (see, for example, Hurley, 2024; OxfordHouse, 2018; Shemesh, 2022), and many influential language teachers on websites like YouTube provide free shadowing exercises (see, for example, English like a Native, 2018; English with Lucy, 2023; Pronunciation with Emma, 2022).

Whilst the technique emphasises simultaneity, in practice shadowing involves a short lag between the audio stimulus and its repetition. Hamada (2017) illustrates this using following example:

Shadowing:
CD: Akita is located in the Tohoku region. It is famous for rice…
Learners: Akita is located in the Tohoku region. It is famous for rice…
(p. 6).

There are also several variations on the standard shadowing technique described above. Hamada and Suzuki (2022), for instance, list the following 16 primary types of shadowing:

1) Standard shadowing.
2) Mumbling (which involves shadowing quietly).
3) Text-presented shadowing (which encompasses scripts).
4) Pre-shadowing (shadowing before learning the content of an audio fragment).

5) Post-shadowing (shadowing after learning the content of an audio fragment).
6) Self-monitoring shadowing (shadowing whilst recording oneself, then reviewing the recording).
7) Pair-monitor shadowing (shadowing with a pair who monitors).
8) Prosody shadowing (shadowing giving special attention to prosody).
9) Gesture shadowing (shadowing using gestures to signal the strength or weakness of words).
10) IPA shadowing (shadowing with a transcription written in the International Phonetic Alphabet).
11) Content shadowing (shadowing with additional focus on the content of the audio).
12) Conversational shadowing (shadowing in pairs).
13) Selective shadowing (shadowing only particular types of words).
14) Phrase shadowing (shadowing one phrase at a time).
15) Shadow reading (shadowing whilst reading, then summarising and retelling).

When working with shadowing in the classroom, different variations may be used in conjunction with each other (Hamada, 2021).

2.2.2. Shadowing vs. related techniques

As noted above, shadowing involves listening to and repeating an audio stimulus as simultaneously as possible. This focus on simultaneity makes the technique different to classical listen-and-repeat exercises, which involve a more prolonged lag between stimulus and repetition. Hamada (2017) illustrates the difference between shadowing and standard listen-and-repeat in the following way:

Shadowing:
CD: Akita is located in the Tohoku region. It is famous for rice…
Learners: Akita is located in the Tohoku region. It is famous for rice…

Repetition:
CD: Akita is located in the Tohoku region
Learners: Akita is located in the Tohoku region
(p. 6)

This difference is considered important, as shadowing, and more specifically the way in which its online, simultaneous nature limits cognitive resources, is thought to help direct learners' attention to the incoming sounds themselves, rather than their meaning (Hamada, 2016b). Repetition, on the other hand, is considered an offline task, thought to split learners' cognitive resources between sounds and meaning (Hamada, 2016b).

Hamada (2017) also highlights that shadowing differs from a related technique named "mirroring". Mirroring is a project-based technique (LaScotte et al., 2021), which involves students selecting a model speaker and carefully studying and imitating their verbal and non-verbal delivery of a speech over a number of weeks. In mirroring, learners assess the purpose of a speech segment, before analysing pronunciation features, like intonation, stress and pausing, and speakers' gestures. After analysis, learners practise copying the speech, staying as close as they can to the delivery of the model speaker. However, unlike shadowing, listening and pronunciation practice are not necessarily simultaneous. After various weeks of practice and feedback, mirroring culminates in a performance of the speech in class and may also involve learners then preparing their own speech and attempting to "channel" the voice of their chosen speaker (LaScotte & Tarone, 2022).

Whilst listen-and-repeat and mirroring may also be valuable pronunciation training exercises worth exploring in more depth, research into shadowing for this systematic review was considered more pertinent for several reasons. On the one hand, listen-and-repeat is a well-established, traditional technique well studied in the research literature (Jones, 1997), making a systematic review into its effects of less academic interest. On the other hand, although mirroring is a novel exercise, at present it has not experienced the same growth as shadowing, and little research on the technique is available. As such, a systematic review of mirroring was not considered as timely or relevant as shadowing.

2.2.3. Empirical research: shadowing for listening

To date, most of the research on shadowing has focused on its application as a tool to train listening skills (Hamada, 2021). Indeed, the technique has been described as "an intensive exercise regime to train the L2 listening muscle" (Hamada, 2017, p. xiv), and a range of classroom studies indicate that shadowing can help improve learners' overall listening comprehension (Kato, 2009; Lin, 2009, Mochizuki, 2006; Tamai, 1997). More specifically,

the technique can help improve skills like phonemic perception (Hamada, 2016a; Kadota, 2019), phonemic discrimination (Hamada, 2020), and word recognition (Ekayati, 2020). In addition, shadowing may help learners adapt to unfamiliar accents in an L2 (Hamada, 2019b), with script-based shadowing being particularly effective for this purpose (Hamada & Suzuki, 2020).

The degree of improvement in the abovementioned listening sub-skills appears to depend on the exact shadowing procedure followed (Hamada, 2022). Furthermore, for shadowing to be most effective for training listening skills, Hamada (2018a, 2021) recommends following a series of principles, which include:

- Selecting an audio text to match students' goals in terms of delivery and speed.
- Using shadowing after learning the content of an audio text, to allow for greater concentration on phonological features.
- Practising each audio text for five rounds. Above this, student improvement in number of correctly shadowed syllables seems to plateau (Shiki et al., 2010).
- Giving students a sense of purpose and improvement by clarifying the objective of the exercise, and incorporating recordings and review stages to talk with students about their progress.

2.2.4. Potential mechanisms behind shadowing and listening improvement

Proponents of shadowing claim that it is more effective than classic listening or listen-and-repeat exercises in training bottom-up processing for listening. That is, processing of the smaller units of an acoustic message, like individual sounds and phonemes.

This is because, in standard listening exercises, learner attention tends to be focused on comprehension. For example, when listening to an audio track and answering questions about it, learners naturally pay the most attention to meaning to be able to find the information they need. If these learners have a low listening proficiency, they may continually rely on top-down processing, in the form of background knowledge, extralinguistic information, prediction or note taking (Hamada, 2018a), to compensate in such exercises. This leaves their bottom-up listening skills underdeveloped (Hamada, 2017).

In shadowing, on the other hand, learners are encouraged, both by teachers and by the online nature of the technique, to pay most attention to the recognition and vocalisation of phonemes, rather than meaning (Kadota, 2019). This directly develops bottom-up skills (Hamada, 2017). In addition, when compared to classical listen-and-repeat exercises, the short lag between stimulus and repetition in shadowing is thought to "block" learners from engaging deeply with meaning and direct their attention more strongly to the incoming sounds (Hamada, 2018a).

2.3. Gaps in the literature

Although there is a substantial body of literature exploring shadowing and listening, research into shadowing as a pronunciation training technique is more limited (Foote, 2017; Hamada, 2019a). In theory, shadowing could help improve pronunciation by both encouraging students to pay close attention to features of speech like phonemes, stress, and intonation, and by training muscle memory through the almost simultaneous repetition of these speech features (Hamada, 2018a, p. 22). It also may help train conversational speaking skills by mirroring the process of listening to input speech and producing an almost immediate response (Kadota, 2019). However, given how cognitively challenging shadowing is, researchers caution that attempting to develop listening and pronunciation simultaneously may be too challenging for learners and thus miss achieving either objective (Hamada, 2018a, p. 20). Using shadowing for pronunciation may, therefore, be more suitable for more advanced learners who already possess good phoneme perception skills (Hamada & Suzuki, 2022). More research into these questions, and into shadowing for speaking in general, has been recommended (Hamada, 2021).

Another under-researched area in the field of shadowing is learner attitudes to the technique (Hamada, 2021). This is particularly important given that, as a practice, shadowing can be very psychologically demanding and repetitive, which can lower learners' motivation (Hamada, 2021). Student attitudes to repetitive practice can also have a cultural component, with some "cultures of learning" employing, and normalising, such techniques more than others (Cortazzi & Jin, 1996). For example, whilst in particular national contexts, such as China, students may see repetition as a valuable practice (Hu, 2002), repetitive exercises may be less compatible with current Western paradigms of communicative language teaching. In addition, shadowing is currently popular in Asia and, as such, much research to date has been

conducted with Asian learners (Kadota, 2019). It is therefore essential to explore student responses to the practice in a wider range of national settings and cultures of teaching and learning (Hamada, 2018a, p. 21).

Finally, research in shadowing has often been carried out with English and Japanese learners (Kadota, 2019). More research into use of shadowing to teach other languages, like Zajdler's (2020) work on shadowing and the Chinese language, is therefore needed.

2.4. Research questions

Given calls to explore both shadowing for pronunciation and student attitudes to the technique, the following overarching research questions was developed to guide this systematic review:

What is the extent and nature of research on the use of shadowing in teaching pronunciation in L2 classrooms?

This research question was broken down into the following sub-research questions:

RQ1. What evidence is available on the effectiveness of shadowing to improve learners' pronunciation?

RQ2. How do learners evaluate this teaching and learning technique?

3. Methodology

This section provides an introduction to systematic review methodology, and justifies its appropriateness for the current study. It then outlines the key tools used and procedures followed when conducting the review.

3.1. Systematic reviews: principles, application, and appropriateness

Systematic reviews provide an overview of existing knowledge on a topic, yet differ from traditional narrative reviews in their commitment to rigour and transparency in each step of the research process (Zawacki-Richter et al. 2020). Indeed, systematic reviews are often defined with reference to their "explicit, accountable rigorous research methods" (Gough et al., 2017, p. 4). According to Macaro et al. (2012, p.3), these methods include five key features or principles:

1. Transparency in procedures used throughout the research process, including an initial protocol specifying how the review will be conducted.
2. The inclusion of studies based on exhaustive, reliable searching.
3. The aim of reducing reviewer bias as far as possible.
4. The aim of producing syntheses with clear messages about the reliability of the evidence reviewed.
5. The aim of ensuring that research is relevant and accessible to end users.

As such, systematic reviews are useful as a tool to provide an overview of a research field in which author bias has been minimised as much as possible (Macaro, 2019). They are also important as most research can only be understood in context: few single studies are so generalisable and methodologically sound that they can present a good approximation of the overall knowledge about a given intervention or research topic (Petticrew & Roberts, 2006, p. 2-3). Furthermore, in the fields of education and applied linguistics, systematic reviews can summarise key research findings in a condensed format, enabling policymakers and stakeholders to make informed decisions (Macaro, 2019; Zawacki-Richter et al., 2020).

Systematic reviews are not always considered appropriate as a research method. However, they are recommended in situations in which there is uncertainty about the effectiveness of an

intervention, or in which some key questions about an existing body of research remain unanswered (Petticrew & Roberts, 2006, p. 21).

Current research into shadowing for pronunciation meets the above criteria. Firstly, as highlighted in section 2.3 (p. 17), there is a wide range of research into shadowing in general, but predominantly in the field of shadowing for listening. Researchers have highlighted that the effectiveness of shadowing for pronunciation, and students' responses to the instructional technique, remain underexplored (Hamada, 2019a; 2021). Secondly, as outlined in section 2.2.1 (p. 13) shadowing has also experienced a rapid growth in popularity in recent years, making it an appropriate moment to consolidate knowledge and fill research gaps surrounding the technique. For these reasons, a systematic review was considered an appropriate research method to explore the intervention.

3.2. The present review

3.2.1. Protocol

In May of 2023, a protocol document (provided in Annex 1) was developed to guide the review, and ensure transparency, consistency, and accountability throughout the research process. PRISMA-P guidelines (Moher et al., 2015) were used to develop the protocol. The items from the guidelines judged most relevant to an MSc level, qualitative synthesis review were included, with others excluded (for example, items relevant to meta-analysis and quantitative exploration of aggregated data) or modified (for example, risk of bias was subsumed in a category about quality assessment of studies).

The protocol included administrative information about the review (for example its title, author, and contributors), and methodological information (such as the rationale, objectives, eligibility criteria, and information sources of the review). It also included basic information on the processes of study selection, data collection, and data synthesis. The protocol was updated in June 2023 to involve search strings, once these had been created.

3.2.2. Eligibility criteria

Before beginning the search for articles, four eligibility criteria were established for studies. Table 1 shows each criterion, and their rationale.

Table 1. *Eligibility criteria*

Type of eligibility criteria	Inclusion criteria	Exclusion criteria	Rationale
Empirical evidence	The studies provided empirical evidence.	The studies did not provide empirical evidence.	The purpose of the review was to evaluate research base.
Year of publication	The studies were published after 1992.	The studies were published before 1992.	As highlighted in the literature review, the first documented case of shadowing applied to L2 language learning was Tamai (1992). This was taken as a logical starting point for research into shadowing for pronunciation.
Topic: L2 learning	The studies investigated shadowing for L2 language learning.	The studies did not investigate shadowing for L2 language learning.	An initial search indicated a large number of studies in which techniques also named shadowing were applied to fields ranging from L1 speech therapy to interpretation, to medicine, dentistry and beyond. This criterion limited the search to studies relevant to L2 language learning.
Topic: improving pronunciation	The studies investigated shadowing for improving L2 learners' pronunciation.	The studies investigated shadowing for L2 language learning, but not improving pronunciation.	This criterion was considered important to highlight studies in which shadowing was used to improve listening skills, but not pronunciation proficiency.

Participants' characteristics like age, gender, proficiency, L1, and L2 were not restricted, as research with participants of all characteristics was deemed of interest to this review. Both

quantitative and qualitative studies were considered of equal importance to the review, and type of study was also not restricted in this regard.

3.2.3. Selection of databases

Web of Science and Scopus were selected for the literature search due to their wide coverage of academic journals and articles. To provide access to journals in the field of education that may not have been indexed in Web of Science and Scopus, the British Education Index, the Australian Education Index, and the Education Resources Information Centre (ERIC) were also selected. Finally, as grey literature is also of importance in social science research (Petticrew & Roberts, 2006), the database ProQuest Social Science Premium Collection was selected to provide access to unpublished dissertations and theses.

Whilst the databases were selected to give the most extensive coverage possible of the research literature, it should be noted that they did not cover all research available. A search using Google Scholar, for instance, revealed several relevant studies that did not appear in the databases selected. However, Google Scholar was not used as a database in the current study due to workload constraints of an MSc level research project, and the fact that the database has been described as a "poor choice" for systematic reviews due to its constantly changing content, algorithms, and database structure (Giustini & Boulos, 2013). Although the search was unable to encompass all existing research, careful selection and searching of targeted databases was designed to ensure transparent, systematic selection of studies.

3.2.4. Search strings

Search strings were developed with the help of an expert librarian at the Department of Education, with tentative search strings piloted on the 5th of June 2023 and the final search carried out on the 15th of June 2023. The search was designed to balance sensitivity, the inclusion of relevant studies, and specificity, the exclusion of irrelevant studies (Petticrew & Roberts, 2006).

As the aim of the search was to find research on shadowing for training pronunciation, it was considered essential for articles to have both "shadowing" and terms related to oral language ability in the abstract. These terms were very general, such as "pronunciation" or "speaking",

as piloting with more specific terms related to pronunciation (like “intonation”, “prosody”, “segmental”, or “suprasegmental”) did not increase the number of articles located in the search.

To increase the specificity of the search, the abstract of the articles being searched also had to contain terms related to the field of second language learning. Additional filters, for example “NOT medicine” were also added to further reduce irrelevant papers from other fields, which were concentrated in the fields of medicine, dentistry and the study of aerial vehicles. These filters were not used in databases which indexed educational studies only.

The search terms used are shown below:

Table 2. *Search terms used (general)*

	Shadowing	Pronunciation	Second language learning	Filters
Place in article	Abstract	Abstract	Abstract	All document
Term(s)	Shadowing	AND Pronunciation OR Speaking OR Oral OR Speech	AND Learn* OR Stud* OR Language*	NOT Medicine OR Dentistry OR Nursing OR Ultrasound OR Cocktail OR Satellite

The exact search strings were as followed in each database:

Table 3. *Search terms used per database*

Database	Scopus	Web of Science	ProQuest Social Science Premium Collection	ERIC
Search string	(ABS (shadowing)) AND (ABS (speaking OR pronunciation OR oral OR speech)) AND (ABS (learn* OR stud*)) OR (ABS (language*)) AND NOT (ALL (medical OR dentistry OR nursing OR satellite OR ultrasound OR cocktail)	(AB=(shadowing)) AND (AB=(speaking OR pronunciation OR oral OR speech)) AND (AB=(learn* OR stud*) OR (AB=language*)) NOT (ALL=(medical OR dentistry OR nursing OR satellite OR ultrasound OR cocktail))	(abstract("shadowing")) AND (abstract(speaking) OR abstract(pronunciation) OR abstract(oral) OR abstract(speech)) AND (abstract(learn*) OR abstract(stud*) OR abstract (language*)) NOT (all (medical) OR (nursing) OR (cocktail) OR (satellite))	(abstract:shadowing) AND (abstract:pronunciation OR abstract:speaking OR abstract:oral) AND (abstract:stud* OR abstract:learn* OR abstract:language*)
Number of hits	167	333	182	18

The search produced an initial sample of 700 articles. The 682 articles from the databases Scopus, Web of Science, and ProQuest Social Science Premium Collection were exported and uploaded to Rayyan on the 26th of June 2023. As the 18 articles from ERIC could not be

exported in a format compatible with Rayyan, the titles and abstracts were stored in an Excel file.

3.2.5. Duplicates

The articles exported to Rayyan were screened for duplicates, with the system automatically detecting 196 duplicates, and an additional four duplicates detected by hand. The articles stored in Excel format were also screened by hand for duplicates, with a total of 15 duplicates located. All 215 duplicates were removed, leaving a total of 485 articles.

3.2.6. Title and abstract screening

Title and abstract screening took place from the 26th of June 2023- 30th June 2023. The exclusion criteria for this screening process, as defined in the section "Eligibility criteria" were as follows:

- **Exclusion criteria 1:** the articles did not provide empirical evidence.
- **Exclusion criteria 2:** the articles did not investigate shadowing for L2 language learning.
- **Exclusion criteria 3:** the articles investigated shadowing for language learning, but not for improving learners' pronunciation.
- **Exclusion criteria 4:** the articles were published before 1992.

Title and abstract screening was over-inclusive. That is, if the title and abstract did not include sufficient information to determine whether the article met the inclusion criteria, the article was automatically included to move forward to full text screening. In total, 419 articles were excluded, and 66 were included for full text screening. The table below shows the number of articles excluded per exclusion criteria.

Table 4. *Numbers of articles excluded per exclusion criterion*

	Exclusion criterion 1: no empirical evidence	Exclusion criterion 2: not shadowing for language learning	Exclusion criterion 3: not shadowing for pronunciation in language learning	Exclusion criterion 4: published before 1992
Number of articles excluded	1	406	8	4
Total number of articles excluded	419			

Note. All four studies published before 1992 were also not related to shadowing for language learning.

3.2.7. Title and abstract screening by second reviewer

A second reviewer, a fellow student on the MSc Applied Linguistics for Language Teaching, also conducted title and abstract screening for 10% of the sample, excluding duplicates (48 studies). Before the screening, a meeting was held to present the project to the second reviewer and familiarise him with the research questions and exclusion criteria. He was also provided with a document summarising this information to guide his screening. Interrater reliability was calculated after his selection, and agreement was substantial (κ = .622*)*. Disagreements on inclusion were mostly, however, regarding laboratory studies related to phonetic convergence across dialect boundaries after shadowing, either comparing convergence between native speaking participants of different dialects or between native and non-native speakers. The first reviewer had marked these as "exclude", but the second reviewer had marked them as "maybe". After discussion, it was decided that these studies did not directly focus on shadowing for language learning or pronunciation, and therefore did not meet the inclusion criteria.

3.2.8. Full text screening

Full text screening took place between the 30th of June and the 30th of July 2023. During this process, an additional exclusion criterion was added to the four used in title and abstract screening. The additional criterion was the full text of the article not being available in English. This was because one article had an abstract available in English, but the full text in French. Full text screening resulted in the inclusion of a total of 31 articles.

3.2.9. Additional searching

To supplement the electronic searches, hand searching of backward and forward citations was used. Backward citation checking involves screening the reference lists of included articles, then subjecting relevant titles to abstract and full text screening. Forward citation checking, on the other hand, is the screening of articles citing the studies selected for inclusion in the review.

If used alone, backward and forward citation checking could be problematic for a systematic review, due to citation bias (see Gøtzsche, 2022, for a discussion of citation bias). However, when used in combination with electronic searches, systematic review guidelines recommend both forms of citation checking as part of a balanced search strategy (Atkinson et al., 2015; Newman & Gough, 2020; Petticrew & Roberts, 2006).

3.2.10. Backward citation checking

Backward citation checking took place between the 21st and 25th of August 2023. The reference lists of the 31 included articles were screened, with relevant titles then also subjected to full text screening with the same exclusion criteria detailed above.

This resulted in the inclusion of an additional 12 articles, of which five were dissertations or theses, four were conference proceedings, and four were academic articles. The fact that the majority of these additional articles were unpublished or "grey" literature may explain why they did not appear in the electronic searches. One identified dissertation was not available anywhere online and was sourced by contacting the author via ResearchGate.

It should be noted that backward citation also revealed a significant body of research available in Japanese. Nine articles referenced had titles relevant to the research question but were only available in Japanese.

In addition, there were two dissertations of theses with relevant titles that were unavailable in university level, national, or international repositories and four articles for which the journal, issue, or book could not be located.

3.2.11. Forward citation checking

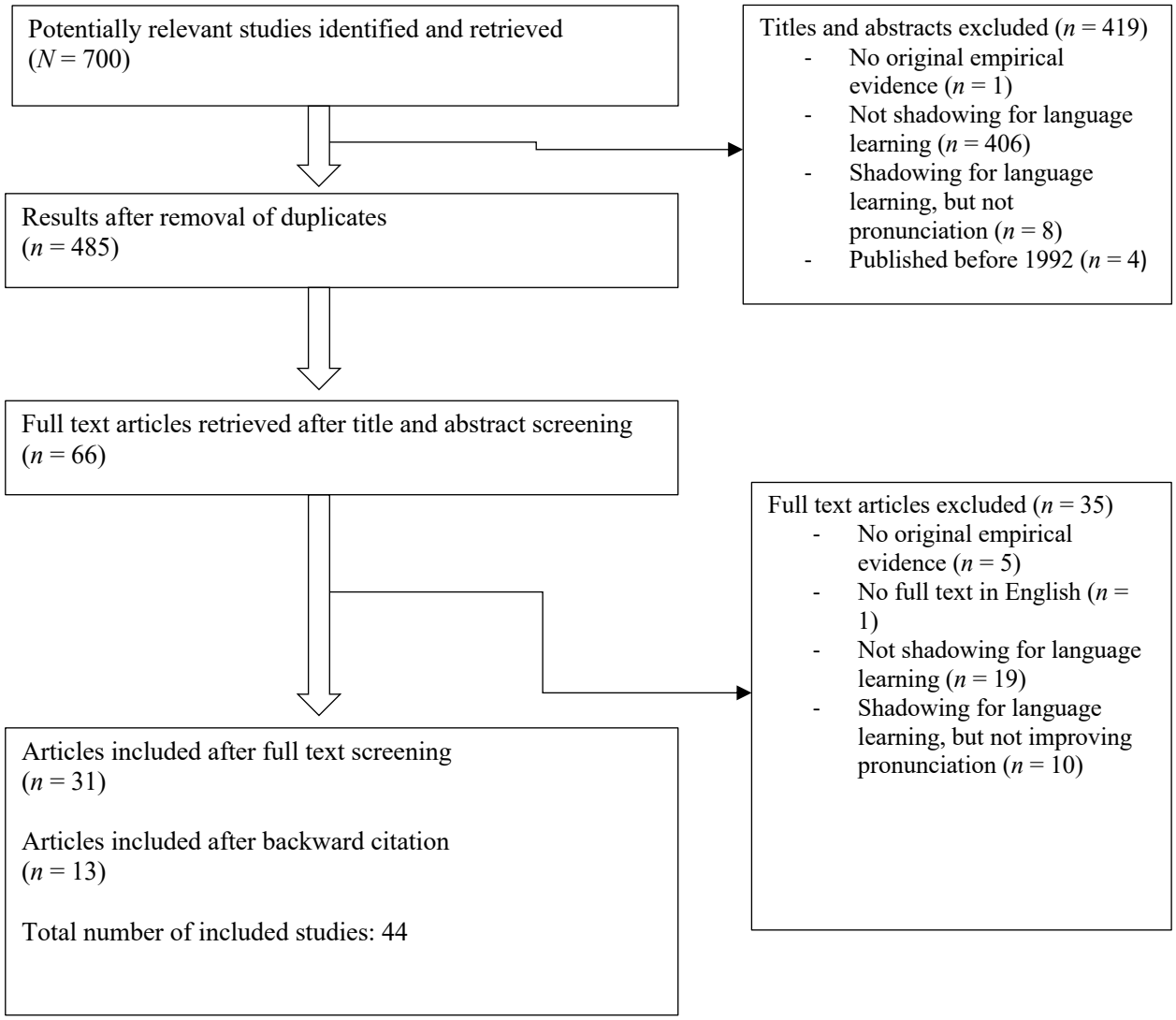
Forward citation checking took place between the 28th of August and the 1st of September 2023. To search for studies citing the 31 included articles, citation tracking functions were used in the following databases: Web of Science, Scopus, and ProQuest Social Sciences Premium Collection. ERIC was not used for forward citation checking as it does not support this function, and the British Education Index and Australian Education Index were not used as no articles were retrieved from the databases in the original search.

Each of the 31 included articles was located in each database, and the databases automatically identified studies citing each article. The titles and abstracts of these studies were then screened. However, bar articles already identified in the original search, no other titles and abstracts met the inclusion criteria. As such, no full texts were screened, and no further articles were included.

3.2.12. Overview of study selection

Figure 1, below, gives an overview of the full process of study selection, and the numbers of studies included and excluded in each step of the process.

Figure 1. *Flow chart of study selection*



3.2.13. Data extraction

Data extraction from the 44 included studies took place between the 15th of January and the 1st of March 2024. It involved systematically compiling key information on each study in a data extraction grid (provided in Annex 2).

The extraction grid included sections on each study's abstract, introduction, literature review, research questions, methodology, sampling, data collection, data analysis, and its results, discussion, and conclusion. The grid also included sections to evaluate each article's weight of evidence according to 1) relevance 2) appropriateness 3) contribution 4) trustworthiness. Ratings of high, medium, and low were used for each criterion. This extraction grid was based on models used in numerous other systematic reviews in applied linguistics studies (i.e., Macaro et al., 2018; Rose et al., 2018; 2021).

The grid also included several questions particular to shadowing and pronunciation research. Firstly, whether the tasks used in the study were in line with shadowing as defined in the literature review. That is, whether tasks involved simultaneous speaking and listening, with participants repeating one chunk of an utterance whilst listening to incoming chunks. Secondly, the task type (controlled or spontaneous) used to collect speech samples, and the scoring method (human raters or acoustic measurements) used to assess pronunciation. These two parameters are considered essential in conceptualising and categorising pronunciation studies (Saito & Plonsky, 2019). Finally, the grid included a section on whether studies could be considered to be in line with the Nativeness Principle or the Intelligibility Principle. To determine this, Thomson and Derwing's (2015) categorisation procedure based on method of pronunciation assessment was used. In this procedure, if speech was assessed in comparison with a native-like target (e.g. Voice Onset Time, pitch contours, error counts, or accent ratings), studies were categorised as Nativeness studies. If speech was assessed using measures of comprehensibility (e.g. ratings) or intelligibility (e.g. transcriptions), studies were categorised as Intelligibility studies.

Based on all this data, a critical review of each study was written and added to the grid. All critical reviews can be consulted in Annex 3.

3.2.14. Narrative synthesis

The extracted data was used to create a narrative synthesis of the studies. First, tables and charts were created to summarise key characteristics of the final selection of studies (e.g., publication date, country, L1, L2, etc.). Next, the results from studies were synthesised per research question. For RQ1, results were organised around the type of pronunciation improvement measured (e.g., global, suprasegmental, segmental). For RQ2, results were organised into perceived enjoyment and interest of shadowing, perceived effectiveness of shadowing, and changes in perceptions over time. The narrative synthesis is presented in the following section on results.

4. Results

This chapter outlines results from data extraction. It first justifies why some of the original 44 studies are not included for discussion, and presents key characteristics of the included studies, such as date of publication, countries, participant characteristics, and research design. More specific information related to pronunciation research, such as speech task type employed, scoring method used, and alignment with Nativeness or Intelligibility principles, is also provided. Next, the chapter present results by research question, with the question of whether shadowing can improve L2 pronunciation addressed first, and student perceptions of the intervention second.

4.1. Characteristics of included studies

4.1.1. Number of studies included

Data extraction revealed several issues that resulted in a reduced number of studies being included for discussion in the present chapter. Firstly, although all studies claimed to employ shadowing, after in-depth review it was determined that eight of the studies did not use the technique as defined in the shadowing literature. That is, these eight studies did not employ tasks in which participants repeated part of an utterance with as little lag as possible, whilst simultaneously listening to incoming parts of the utterance. Instead, these studies used delayed repetition of isolated words or characters.

Given that these studies cannot be considered to explore shadowing as defined in this review, their results are not included in the section that follows. They are, however, included in the larger sample, as it was deemed important to acknowledge that several studies exist in the current literature that claim to investigate shadowing, but do not adhere to widely accepted definitions of what shadowing entails. This has later implications for a discussion on how the term shadowing has been misapplied in the field. Critical reviews for these studies can be consulted in Annex 3, and the studies were the following:

Table 5. *Studies not in line with definitions of shadowing*

Studies not in line with definition of shadowing	Number
Althubyani (2021), Hashimoto et al (2022), Hutchinson (2022), Nakayama (2021), Rojczyk (2013), Šturm et al. (2022), Zając and Rojczyk (2014), Zhang and Peng (2017)	8

Secondly, there were two dissertations whose results were also presented in academic papers several years later, as shown in Table 6.

Table 6. *Results in dissertations and subsequent publications*

Dissertation	Article
Sumiyoshi (2014)	Sumiyoshi and Svetanant (2017)
Willardson (2014)	Martinsen et al. (2017)

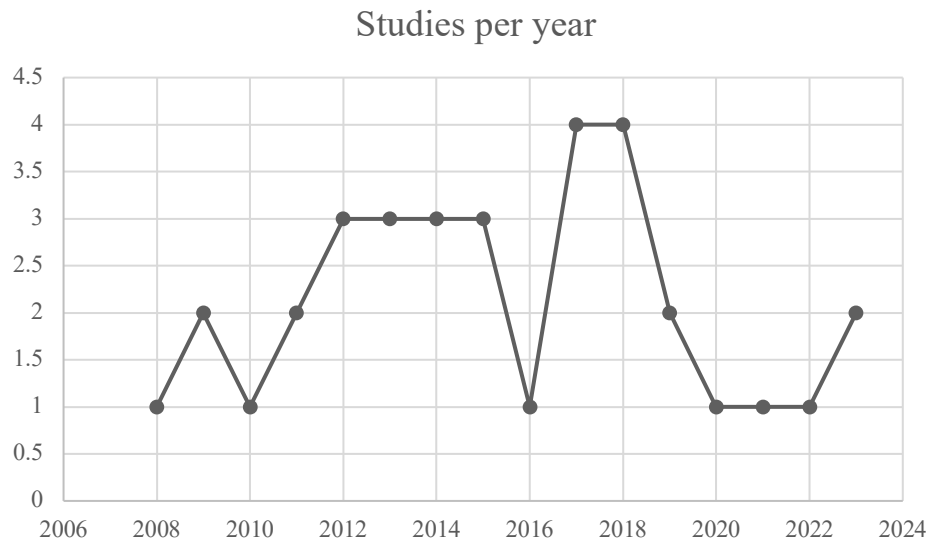
In these cases, the results from the two studies were subsumed into one. The dissertations are referred to in results, as they were both the first published and are the more comprehensive of the two versions of the results.

Overall, the abovementioned issues mean that a total of 34 studies, rather than the original 44, are described in the sections that follow.

4.1.2. Basic information

The included studies were published between 2008 and 2023, with an overall increase in publications between 2010 and 2018, which then dropped in 2019 (see Figure 2).

Figure 2. *Studies by year*



As Table 7 shows, a large number of studies were conducted in Asia (n = 20), particularly in Japan (n = 10) and Taiwan (n = 6). There were also a number of studies conducted in North America (n = 4), the Middle East (n = 2), and Australia (n = 1). It should be noted that seven studies did not report the country of research clearly.

Table 7. *Studies per country*

Country	Number of studies
Australia	1
Canada	2
China	2
Iran	1
Japan	10
Malaysia	1
Mongolia	1
Saudi Arabia	1
Taiwan	6
USA	2
Not stated clearly	7

Most studies were conducted in university settings (n = 18), with two conducted in secondary settings, two in primary settings, two in afterschool centres for primary school students, and two in language schools (see Table 8).

Table 8. *Studies by educational setting*

Educational setting	Number of studies
University	18
Secondary school or equivalent	5
Primary school or equivalent	2
Afterschool learning centre for primary school children	2
Language schools (outside HE)	2
Other	1
No information provided	4

The majority of studies were conducted with adults (n = 21), defined as participants aged 18 and over, and including university students of unspecified age group (see Table 9).

Table 9. *Studies by age group*

Age range	Number of studies
Children >13	3
Young adults 14-18	8
Adults (18+ and university students of unspecified age)	21
Age unstated	3

Note. Studies in which the age ranged from below 18 to the age of 18 were classified as young adults, whereas studies in which the age ranged from 18 to ages above this were classified as adults. Yavari and Shafiee's (2018) study was counted twice, as participants

were aged between 15-20, with no information provided on the exact numbers of participants from each age.

As is shown in Table 10, a large number of studies (n = 17) did not report participants' L1 clearly. In studies that did report L1, the most common were Japanese (n = 7), Mandarin (n = 3), or mixed L1s (n = 4).

Table 10. *Studies by L1*

L1	Number of studies
English	1
Mixed	4
Japanese	7
Mandarin	3
Persian	1
Vietnamese	1
Not clearly stated	17

The majority of studies involved participants learning English as an L2 (n = 27), with some studies focusing on Japanese learners (n = 5), and one focusing on French learners (see Table 11). One study did not report participants' L2 clearly.

Table 11. *Studies by L2*

L2	Number of studies
English	27
French	1
Japanese	5
Not clearly stated	1

The most represented L2 proficiency level in the sample was intermediate (n = 9), followed by elementary (n = 5), and mixed groups of beginner and intermediate learners (n = 3) (see Table 12).

Table 12. *Studies by proficiency*

Proficiency	Number of studies
No prior language experience	1
Elementary	5
Intermediate	9
Advanced	1
Mixed (beginner, intermediate, advanced)	1
Mixed (intermediate and advanced)	1
Mixed (beginner and intermediate)	3
Not stated or measured	13

15 studies had a quasi-experimental design, involving a control or comparison group, and 15 involved classroom interventions (see Table 13). In addition, three studies were user studies of shadowing applications or software, one study was an action research project, and one was a case study.

Table 13. *Type of study design I*

Type of study	Number of studies
Quasi-experimental	15
Classroom intervention	15
Action research project	1
Case study	1
User study	3

Note. Sumiyoshi (2014) appears twice, as in her thesis there are three different studies (two classroom intervention studies, one quasi-experimental study).

As shown in Table 14, 31 studies were longitudinal, and four were cross-sectional.

Table 14. *Type of study design II*

Type of study	Number of studies
Longitudinal	31
Cross-sectional	4

Note. Sumiyoshi (2014) appears twice as two of the studies were longitudinal, and one was cross-sectional.

As Table 15 illustrates, the most common type of data collection instruments in the studies were different types of pronunciation tests (n = 31) and questionnaires (n = 16). Some studies also employed interviews (n = 6), student logs and diaries (n = 2), and teacher observations (n = 5).

Table 15. *Types of data collection instruments*

Type of data collection instrument	Number of studies
Pronunciation tests	31
Questionnaire	16
Structured interviews	1
Semi-structured interviews	4
Interviews (unspecified type)	1
Student logs or journals	2
Teacher/researcher observations	5

4.1.3. Pronunciation-related information

The following tables and data are related only to the 31 studies involving measures of pronunciation. In Saito and Plonsky's (2019) influential framework for conceptualising pronunciation studies, two key aspects are task type (controlled vs. spontaneous) and type of scoring method used (human raters vs. acoustic analyses).

In the sample, as shown in Table 16, there was a predominance of controlled tasks, with 20 studies employing controlled tasks only. Four studies used only spontaneous tasks, and four studies used a combination of controlled and spontaneous tasks. In a further four studies, the task used was unclear.

Table 16. *Types of speech elicitation task*

Type of task	Number of studies
Only controlled	20
Only spontaneous	3
Both	4
Unclear	4

As illustrated in Table 17, 12 studies employed only human raters, five studies used solely acoustic analyses or computer programs, and five studies used a combination of the two. Nine studies did not report scoring method clearly.

Table 17. *Type of scoring method used*

Type of scoring method	Number of studies
Human raters	12
Acoustic analyses or computer programs	5
Both	5
Unclear	9

Another key aspect of research into pronunciation is whether studies align with the Nativeness or Intelligibility principles (Thomson & Derwing, 2015). As Table 18 shows, based on their method of pronunciation assessment, 12 studies were in line with the Nativeness Principle, and two were in line with the Intelligibility Principle. Four studies showed elements of both principles, and for 13 studies it was impossible to determine based on the information given, or irrelevant given the measurement used (e.g., fluency measured by words spoken per minute).

Table 18. *Alignment with Nativeness or Intelligibility Principles*

In line with	Number of studies
Nativeness Principle	12
Intelligiblity Principle	2
Elements of both	4
Unclear or neither	13

4.2. What evidence is available on the effectiveness of shadowing to improve learners' pronunciation?

A total of 26 studies directly explored how shadowing can improve learner pronunciation, and are explored in detail in this section. One additional study focused on a comparison of different types of shadowing (Hamada, 2018b), and four articles detailed user studies of different applications and software for shadowing (Kurniawan, 2019; Nakanishi & Nakanishi, 2015; Zhang et al., 2016; 2020). However, due to space constraints, these studies are not described in the present section as they were deemed to be less directly relevant to the research question. In addition, the user studies were considered low quality, fragmented research. Critical reviews for these studies can be consulted in Annex 3.

The 26 studies explored focused on a range of different features of pronunciation, and, as such, measured pronunciation improvement in different ways. Results are presented below according to aspect of pronunciation proficiency studied: global/holistic pronunciation control (such as measures of comprehensibility, intelligibility, and accentedness),

suprasegmental control (e.g., fluency, prosody), and segmental control (i.e., production of specific phonemes).

4.2.1. Improvements in global/holistic pronunciation control

Eleven studies explored the impact of shadowing on global or holistic measures of pronunciation. Key information from these studies is shown in Table 19.

Table 19. *Key information holistic studies*

Study	Construct	Task type	Scoring system	Results
Bovee and Stewart (2008)	Accentedness	Controlled	Human raters	Positive
Foote and McDonough (2017)	Comprehensibility and accentedness	Controlled and spontaneous	Human raters	Positive for comprehensibility, negative for accentedness
Guo and Hsu (2017)	Speaking proficiency	Controlled	Human raters	Positive
Hori (2008)	Intelligibility and accentedness	Controlled	Human raters and acoustic analysis	Negative
Huang et al. (2023)	Intelligibility	Controlled	Human raters	Positive
Lin (2009)	Speaking proficiency	Controlled	Human rater	Positive
Mishima and Cheng (2017)	Intelligibility and speaking proficiency	Controlled	Human raters	Positive
Rongna et al. (2013)	Naturalness	Controlled	Human raters and acoustic analysis	Positive
Rongna et al. (2015)	Naturalness	Controlled	Human raters and acoustic analysis	Positive
Shao et al. (2023)	Comprehensibility and accentedness	Controlled and spontaneous	Human raters	Positive
Willardson (2014)	Comprehensibility and accentedness	Controlled and spontaneous	Human raters	Positive

The majority of studies focused on either one or a combination of comprehensibility, intelligibility, and accentedness. Two studies included measures of comprehensibility (Foote

& McDonough, 2017; Shao et al., 2023), two of intelligibility (Huang et al., 2023; Mishima & Cheng, 2017), and five of accentedness (Bovee & Stewart, 2008; Foote & McDonough, 2017; Shao et al., 2023) or a similar construct (Rongna et al., 2013; 2015). Two studies included a measure that combined elements of accentedness and comprehensibility (Willardson, 2014) or accentedness and intelligibility (Hori, 2008). All studies but one (Hori, 2008) reported positive results. In addition, two studies (Guo & Hsu, 2017; Lin, 2009) also explored the impact of shadowing on overall speaking ability.

The weight of evidence of the studies was assessed as follows:

Table 20. *Weight of evidence holistic studies*

Study	Relevance RQ1	Appropriateness RQ1	Contribution RQ1	Trustworthiness RQ1
Bovee and Stewart (2008)	High	Medium	Medium	Medium
Foote and McDonough (2017)	High	High	High	Medium
Guo and Hsu (2017)	Medium	Medium	Low	Low
Hori (2008)	High	Medium	Medium	Medium
Huang et al. (2023)	High	Medium	Medium	Medium
Lin (2009)	Medium	Medium	Low	Medium
Mishima and Cheng (2017)	Medium	Medium	Medium	Medium
Rongna et al. (2013)	High	Low	Medium	Medium
Rongna et al. (2015)	High	Low	Medium	Medium
Shao et al. (2023)	High	High	High	Medium
Willardson (2014)	High	High	Medium	Medium

Guo and Hsu's (2017) study and Lin's (2009) study were considered limited in their relevance, appropriateness, and, most notably, their contribution to RQ1. This is because both studies used a Taiwanese speaking test, the General English Proficiency Test, with a scoring system that encompasses pronunciation, intonation, and fluency, but also content relevance, and usage of grammar and vocabulary (Language Training and Testing Centre, 2016). The composite nature of the scale, and the fact that the authors do not report specific pronunciation-related scores, makes it difficult to disentangle the various components from one another. Whilst these studies both reported positive results, they are therefore not

described in the section that follows, which only focuses on the most salient research. However, critical reviews can be consulted in Annex 3.

4.2.1.1. Results from selected studies

Only two studies (Foote & McDonough, 2017; Shao et al., 2023) were assessed as both highly relevant *and* highly appropriate to RQ1, due to several notable elements of methodological rigour. Firstly, they included both controlled and spontaneous tasks to gain speech samples, which has been deemed essential for ecological validity in pronunciation research (Thomson & Derwing, 2015) and will be discussed in detail in Chapter 5. Secondly, they used multiple raters, whose backgrounds were described in detail, and included calculations of interrater reliability. These features are considered crucial for validity and reliability of pronunciation evaluation (Kang & Kermand, 2017; Lee et al., 2015). The studies also used different scales for different constructs, rather than measuring improvements in comprehensibility and accentedness via a single scale. Finally, both studies reported the shadowing procedures followed in detail.

Foote and McDonough (2017) researched the effect of an 8-week shadowing programme on the pronunciation of 16 advanced ESL learners in Canada, aged 18-83. Pre-, mid-, and post-tests were administered in weeks 1, 6, and 8. The tests included a picture narrative task, developed by Derwing et al. (2009), and a shadowing task. The picture narrative task was used to increase comparability with other research with immigrants. It should be noted this test was the same at pre- and post-test.

Test recordings were assessed by 22 English L1 raters, who were recruited from a Canadian university and trained to use a computer-based rating system on MATLAB. Raters evaluated speech samples from the picture task for accentedness, comprehensibility, and fluency, using one sliding scale per construct. They also rated samples from the shadowing task for how well participants imitated the speech model. Raters heard recordings in a randomised order and were not aware of which time point the recordings were from. Inter-rater reliability was calculated using intraclass correlation coefficients, and a high degree of agreement was found on all measures: shadowing (α=.86), accentedness (α=.91), comprehensibility (α=.89), and fluency (α=.93). Results showed that, overall, participants showed significant improvement in comprehensibility (p=.01), shadowing (p=.0001), and fluency (p=.0001), although not in

accentedness (p=.05). However, despite overall improvement, there was not always improvement from one testing time to the next, indicating potential minimum thresholds for improvement or possible points at which the benefits of shadowing decline.

Overall, this study is methodologically robust and provides evidence to suggest shadowing can lead to ecologically valid improvements in comprehensibility. However, its trustworthiness is limited by attrition of six participants, who may have left the project as shadowing was not effective for them, the use of some identical pre- and post-tests, and the lack of control group. This last point is particularly important as some participants were new arrivals in Canada and L2 pronunciation improvement in adults often takes place in the first year of exposure in naturalistic environments (Flege, 1988; Munro & Derwing, 2008).

Shao et al. (2023) also investigated the effect of shadowing on students' comprehensibility and accentedness. Participants were 47 Chinese high school students, aged 17-18, all of whom had been studying English for around 6 years. An experimental group (n = 37) received 2 weeks of training, including 12 30-minute shadowing sessions. A control group (n = 10) received grammar, vocabulary, and writing training instead of the shadowing practice.

Two pre- and post-tests were administered in weeks 1 and 4. One test involved reading six sentences selected from the shadowing materials, and was identical at pre- and post-test. The other was a picture description task, with two different versions counter-balanced across group and time to minimise the impact of topic. 188 test recordings were analysed by five raters, all of whom were advanced L2 speakers of English and graduate students in Applied Linguistics. Raters were given detailed explanations of the constructs of accentedness and comprehensibility, and rating guidelines. Speech samples were rated on 9-point scales for comprehensibility and accentedness. Inter-rater reliability was quite high (.88 for comprehensibility and .85 for accentedness), and only the first 30 seconds of each sample were used to reduce rater fatigue. It is unclear whether raters were blind as to which time point each sample was from.

Results indicated that the experimental and control groups' performances were comparable for most contexts at pre-test, except for comprehensibility in the controlled task, for which the experimental group had a higher score, which was marginally significant (p=.028). In terms of improvement, the experimental group significantly improved their comprehensibility

scores over time on both the controlled task (p<.001) and the spontaneous task (p<.001). The control group's gains in comprehensibility did not reach statistical significance on either task (p=.693, .134). For accentedness, the experimental group improved in both controlled (p<.001) and spontaneous tasks (p<.001). The control group showed no significant changes in accentedness in either task (p=.053, .168).

Overall, this study provides support for the use of shadowing to improve comprehensibility and accentedness. However, the significant differences in size and initial comprehensibility between the control and experimental groups, as well as the fact that the controlled test included sentences from the experimental group's shadowing materials, somewhat limit the trustworthiness of results.

It should be noted that Willardson (2014) also used controlled and spontaneous tasks (a read aloud task and a picture description task) and multiple raters to measure improvements in comprehensibility and accentedness. After a 10-week shadowing intervention, 19 American high-school learners of French showed a significant improvement in comprehensibility and accentedness on the controlled task (p=0.00005), with lower and higher proficiency students also showing significant improvements on the spontaneous task (p=0.05, p=0.03, respectively). Mid-level proficiency students improved, but not significantly, on the spontaneous task (p=0.16), which influenced the overall significance of improvement on this task (p=0.20). However, results from this study are less meaningful than those of Foote and McDonough (2017) and Shao et al. (2023), as the scale used for rating was a 5-point global scale, in which 1 was heavily accented speech that was difficult to understand and 5 was near-native pronunciation able to be understood "by all". This scale makes it difficult to disentangle the effect of shadowing on the individual constructs of accentedness and comprehensibility. The use of identical pre- and post-tests, and lack of a control group, is also problematic.

4.2.1.2. Results from less appropriate studies

The remaining six studies involving holistic measures of pronunciation improvement were considered less appropriate for RQ1 as they only involved controlled tasks to collect speech samples: memorised presentations in the case of Mishima and Cheng (2017) and Huang et al. (2023), shadowing recordings in the case of Bovee and Stewart (2008), and a read-aloud task

in the case of Hori (2008). Rongna et al.'s studies (2013; 2015) were considered particularly inappropriate due to their use not only of a highly controlled read-aloud task, but for their use of the same passage in testing and all training sessions, allowing for no extrapolation of learning. Detailed critical reviews of these studies can be found in Annex 3. However, overall, their results concurred with those presented above.

Two studies reported improved intelligibility after shadowing, in the case of Huang et al. (2023) amongst a sample of 90 Taiwanese college juniors after 6 weeks of English shadowing training and, in the case of Mishima and Cheng (2017) amongst a sample of five graduate students in a US university after 2 weeks of English shadowing. It should be noted that in this last study, descriptions of shadowing are very vague, and it is unclear whether a procedure in line with the present review was used. In addition, whilst the study claims to have measured intelligibility, it appears to have actually measured comprehensibility.

One study also reported notably more native-like speech after 10 weeks of shadowing in a sample of 400 Japanese university students learning English (Bovee & Stewart, 2008). Rongna et al.'s (2013; 2015) studies both reported increased naturalness after shadowing amongst learners of Japanese. The studies involved 33 intermediate learners from China and Mongolia (Rongna et al. 2013), and 14 intermediate learners, of unspecified nationality and L1 (Rongna et al., 2015), respectively, and definitions of naturalness appear to be in line with definitions of native-like speech.

However, although the majority of studies reported improvement after shadowing, it should be noted that Hori's (2008) quasi-experimental study, conducted with 43 Japanese learners of English, found no improvements in a mixed impressionistic measure of accentedness and intelligibility at post- or delayed post-test.

4.2.2. Improvements in suprasegmental pronunciation control

In terms of suprasegmental improvement in pronunciation, studies focused on two main features: fluency and prosody.

4.2.2.1 Fluency

Eight studies included a measure of fluency in their design, with two studies (Wang, 2018; Yavari & Shafiee, 2018) focusing exclusively on fluency and the remaining six including a fluency-related measure as part of a composite measure of pronunciation improvement. Weight of evidence was assessed as shown in Table 21.

Table 21. *Key information and weight of evidence fluency studies*

Study	Task type	Scoring system	Relevance RQ1	Appropriateness RQ1	Contribution RQ1	Trustworthiness RQ1
Foote and McDonough (2017)	Spontaneous	Human rater	High	Medium	Medium	Medium
Hori (2008)	Controlled	Acoustic analysis	High	Medium	Medium	Medium
Hsieh et al. (2023)	Controlled	Acoustic analysis	High	Medium	Medium	Low
Mori (2011)	Controlled	Acoustic analysis	High	Medium	Medium	Medium
Ono et al. (2012)	Controlled	Human rater	High	Medium	Medium	Medium
Rongna and Hayashi (2012)	Controlled	Acoustic analysis	High	Low	Medium	Medium
Wang (2018)	Spontaneous	Unclear	High	Medium	Medium	Medium
Yavari and Shafiee (2018)	Spontaneous	Human rater	High	Medium	Medium	Medium

Three studies used spontaneous tasks (Foote & McDonough, 2017; Wang, 2018; Yavari & Shafiee, 2018) to gather speech samples to assess for fluency, and two studies used controlled tasks (Hori, 2008; Mori, 2011; Ono et al., 2012; Rongna & Hayashi, 2012). One study did not specify the task type used (Hsieh et al., 2023). The fact that no studies combined both task types, and the fact that those using acoustic analysis did not combine these analyses with listener judgments, limits the appropriateness and contribution of many studies. These issues will be unpacked in detail in Chapter 5.

4.2.2.1.1 Results from spontaneous tasks

Results from spontaneous tasks were all positive, with Yavari and Shafiee's (2018) study illustrative of how the impact of shadowing on fluency was explored via this task type. Purposive sampling, targeting students with a Preliminary English Test score with one standard deviation from the mean, was used to select 60 learners out of a group of 112 intermediate level learners of English in Iran. All were Persian L1 speaking females, aged 15-20. Participants were randomly assigned to one of 4 groups: a shadowing only group, a tracking only group, a shadowing and tracking group, and a control group. Shadowing was defined as listening and speaking along, simultaneously, with audio input, whilst tracking was defined as listening and repeating phrase by phrase. Participants in the experimental groups received two 15-minute shadowing or tracking sessions per week, as part of normal classes, over a 5-week period. Videos used for the different experimental groups were the same. However, it is unclear what the control group received.

An identical pre- and post-test was used, which consisted of a semi-structured interview in which participants spoke about holiday plans and famous buildings in their town. Fluency was measured by two raters, according to syllables spoken per minute, and inter-rater reliability was high (.79 at pre-test, .81 at post-test). There were no statistically significant differences in fluency between groups at pre-test. There was, however, a statistically significant difference between the groups at post-test, with the three experimental groups showing significantly higher mean scores than the control group (p<0.05). In addition, shadowing was significantly more effective than tracking (p=.000), and shadowing and tracking was significantly more effective than both shadowing alone and tracking alone (p=.000). These results are, however, limited by the fact that the procedure used for the control group is unspecified. This is important, as the content of the videos seemed to be

related to the content of the pre- and post-test. As such, if the control group did not watch the videos, not being repeatedly exposed to the same vocabulary and grammatical structures could have resulted in lower fluency scores.

Yavari and Shaviee's (2018) results were echoed in Wang's (2018) study, which also used a spontaneous task at post-test (an impromptu presentation) and measured fluency as words spoken per minute. Participants were 40 sophomore Chinese university learners of English, divided into an experimental (n = 20) and control group (n = 10). An unspecified pre-test revealed no significant differences in fluency between the groups, and both groups received 1 hour of class per week for 10 weeks with the same learning materials, topics, schedule, and teacher. However, the experimental group integrated shadowing into the classes and had more of a focus on oral expression, whilst the control group did not. The experimental group scored significantly higher (p=0.001) on fluency than the control group at post-test, which Wang (2018) interprets as clearly linked to shadowing. However, these results should be taken with caution given the potential confounding variable of focus on oral expression in the experimental group's classes. The lack of information on the pre-test also raises the question of whether fluency gains may have been influenced by topic and task, as well as by shadowing.

Finally, Foote and McDonough (2017) also reported improvements in fluency when speech samples from a spontaneous task were rated by human raters on a sliding scale on MATLAB. The design of the study is described in detail in section 4.2.1, but, in short, 16 L2 English speakers showed significant increases in fluency on a picture description task (p=.0001) after an 8-week shadowing programme. However, it should be noted that use of an identical pre- and post-test could have influenced improvements in fluency in this study.

4.2.1.1.2 Controlled tasks only

The remaining studies explored improvements in fluency through controlled tasks, mostly in the form of read aloud tasks. Mori (2011), for example, conducted a study involving a 10-week shadowing and oral reading intervention with 30 Japanese second-year university students. Identical pre- and post-tests, involving a read-aloud exercise, were used to collect speech samples. Eight native speakers were also recorded reading the same test, to allow for comparison with a controlled model. Results from acoustic analysis indicated that mean

durations of both the whole passage and two selected clauses were significantly shorter in the post-test than the pre-test ($p<0.05$), indicating an increase in speech rate and fluency.

Similar to Mori's (2011) results, Rongna and Hayashi (2012) reported a significant increase in speech rate ($p<0.001$) on a read-aloud task amongst 15 intermediate-level students of Japanese after 7 weeks of shadowing training. However, use of the same text during all shadowing and read-aloud testing seriously limits the ecological validity of these results. Ono et al. (2012) also used read-aloud tasks to measure improvements in pronunciation among 34 second year students of English at a technology college in Japan, after a 4-week shadowing programme. Factor analysis indicated that increased fluency, as measured by impressionistic judgments by an unspecified number of raters, was a key factor contributing to improvements in oral reading scores.

Finally, two studies used different task types and measures of fluency. Hsieh et al. (2023) reported increased fluency after shadowing, as computed by an unspecified measure on a computer program called My English Teacher. This study reported that, amongst a sample of 14 English learners at Taiwan National University, an experimental shadowing group ($n = 7$) significantly outperformed a control group ($n = 7$) ($p=0.001$). However, the lack of description of measurements, tasks, and procedures seriously limits the trustworthiness of results. Hori (2008), on the other hand, explored the effect of shadowing on articulation rate of monosyllabic and polysyllabic words. The study, involving 43 Japanese learners of English, reported improved performance in an experimental group ($n = 26$) compared to a control group ($n = 17$) at post-test, but not delayed post-test (Hori, 2008).

4.2.2.2 Prosody

In total, eleven studies explored the impact of shadowing on different elements of prosody. The task type, scoring method, and weight of evidence for these studies is shown in Table 22.

Table 22. *Key information and weight of evidence prosody studies*

Study	Task type	Scoring method	Relevance RQ1	Appropriateness RQ1	Contribution RQ1	Trustworthiness RQ1
El-Esery (2021)	Controlled	Human rater	High	Medium	Medium	Medium
Hori (2008)	Controlled	Acoustic analysis and human rater	High	Medium	Medium	Medium
Hsieh et al. (2023)	Controlled	Acoustic analysis	High	Medium	Medium	Low
Kunihara et al. (2022)	Controlled	Acoustic analysis	High	Medium	Medium	Medium
Kuo and Chou (2014)	Controlled	Human raters	High	Medium	Medium	Medium
Mori (2011)	Controlled	Acoustic analysis	High	Medium	Medium	Medium
Nakayama and Armstrong (2011)	Controlled	Human raters	High	Medium	Medium	Medium
Nguyen and Nguyen (2019)	Unknown	Unknown	High	Low	Low	Low
Omar and Umehara (2010)	Controlled and spontaneous	Human rater	High	Medium	Medium	Low
Ono et al. (2012)	Controlled	Human rater	High	Medium	Medium	Medium
Rongna and Hayashi (2012)	Controlled	Acoustic analysis	High	Low	Medium	Medium
Sumiyoshi (2014)	Controlled	Acoustic analysis	High	Medium	Medium	Medium

The majority of studies used highly controlled tasks to elicit speech samples, such as shadowing recordings (El-Esery, 2021; Nakayama & Armstrong, 2011; Sumiyoshi, 2014), read-aloud tasks (Hori, 2008; Ono et al. 2012; Mori, 2011; Rongna & Hayashi, 2012), or a combination of the two (Kunihara et al., 2022). Only one study used a more spontaneous task: participants' speaking in class, as assessed by teacher observation (Omar & Umehara, 2010). Two studies did not specify task type (Hsieh et al., 2023, Nguyen & Nguyen, 2019). In addition, most studies used either acoustic analysis or human raters, with only one study (Hori, 2008) combining both methods. These issues limited the appropriateness of many studies, and will be discussed in detail in Chapter 5.

4.2.2.2.1 Human ratings of rhythm, intonation, and stress

Five studies explored overall measures of rhythm and intonation through use of human raters. Kuo and Chou (2014), for example, conducted a quasi-experimental study exploring word and sentence level pronunciation with four intact classes of fourth graders in Taiwan. Classes were randomly and equally assigned to experimental or control groups. However, one control group was later excluded, due to the fact that it contained a large number of students with learning difficulties. This left the number of students in the control (n = 26) and experimental (n = 53) groups unbalanced.

The experimental group received a 12-week intervention in which they practised shadowing for 10 minutes per day, four days per week. The control group received no shadowing training, and instead did English homework. Improvement in pronunciation was measured by a 100-word read-aloud test developed by the researchers by referring to similar tests in previous studies. The test was piloted and revised to develop its validity. Students' reading was scored by two Taiwanese EFL teachers, with scores given for correct pronunciation of words (including correct pronunciation of phonemes and stress) and sentences (including correct pronunciation of intonation and chunking). Interrater reliability was calculated and was very high (0.99 at word level and 0.98 at sentence level). However, there was no mention of whether rating was blind to time point of recordings.

Results showed that there were no significant differences between groups at pre-test (p>.05). However, the experimental group significantly outperformed the control group at post-test on measures of word, sentence, and overall pronunciation (p<.05). In addition, gain scores

showed significant intergroup differences in pronunciation at word level, sentence level, and overall (p<.05), suggesting that the experimental group made significantly more improvement than the control group. Whilst these results are positive, they should be taken with caution due to the imbalance in the two groups.

Two other studies using human raters reported similar results, in terms of overall improvements in intonation and rhythm, although the generalisability and trustworthiness of their results was more limited. Ono et al. (2012) reported that improved stress, volume, and intonation were key factors contributing to read-aloud improvements amongst 34 second year students of English at a technology college in Japan after a 4-week shadowing intervention. Details on participants' proficiency, age, and L1, and number of raters, were not provided. Omar and Umehara (2010) also reported observed and self-reported improvements in English rhythm amongst four retired Japanese learners of English living in Malaysia, after a 6-month shadowing action research project. These conclusions were based on participant observation, student diaries, questionnaires and analysis of shadowing recordings. However, their trustworthiness is limited, as, although the authors explain their methods of analysis for the qualitative data, no excerpts are provided to evidence claims. In addition, no explanation or evidence the analysis used for shadowing recordings is provided.

Finally, two studies reported improvements in production of weak forms after shadowing training. Nakayama and Armstrong (2011) conducted a quasi-experimental study with 95 first-year Japanese university students, involving a 6-week shadowing programme. Participants were divided into a Visual Auditory (VA) group (n = 48), in which participants received both visual and auditory input when shadowing, and a Scaffolded Auditory (SA) group (n = 47), in which shadowing moved through three stages of priming: slow enunciated input, careful pronunciation, and relaxed pronunciation. All participants shadowed the same 274-word speech as a pre-test, post-test, and during training. The two authors then counted the number of words correctly shadowed in each recording. Results indicated that, although the overall main effect for both groups was not significant (p>.05), there was a significant effect for function words (p<.05). In addition, the VA group outperformed the SA group in the post-test (p<.01). Whilst this study does provide evidence that shadowing, in particular VA shadowing, can improve participants' production of weak forms, its ecological validity is highly limited due to use of the same highly controlled task for all testing and training. Nevertheless, its results do concur with those of El-Esery (2021), who reported improved

shadowing and recognition of weak forms amongst 35 English learners at a university in Saudi Arabia after a 3-month shadowing intervention. It should be noted that several key details of this last study, such as the shadowing intervention used, and number of raters scoring a shadowing-listening test, are not reported. Full reviews of these studies can be consulted in Annex 3.

4.2.2.2.2. Acoustic measures of intensity, duration, and pitch

Four studies explored improvements in suprasegmental features of pronunciation through acoustic measurements of pitch, duration, and intensity.

Mori (2011) conducted a study involving a 10-week shadowing and oral reading intervention with 30 Japanese second-year university students. Identical pre- and post-tests, involving a read-aloud exercise, were used to measure improvements in prosody. Eight native speakers were also recorded reading the same test, to allow for comparison with a controlled model. Acoustic analysis was used to analyse the recordings from participants and native speakers, with two clauses selected for analysis due to their alternance of stressed and unstressed syllables.

Results indicated changes in intensity, duration, and pitch from pre- to post-test. Firstly, the mean intensity of syllables was significantly greater in the post-test than the pre-test, with almost all syllables in the clauses read with significantly greater intensity at post-test ($p<0.01$). In some parts of the two clauses, increased intensity also contributed to a sharper intensity contrast between stressed and unstressed syllables. Secondly, selected final syllables showed a significant increase in duration ($p<0.05$), indicating a move towards lengthening phenomena characteristic of spoken English. In addition, some unstressed syllables (such as "I" and "ing") showed a significant shortening ($p<0.05$) at post-test, indicating a better control of alternate stressed and unstressed syllables after training. Pitch also fell more sharply on some final content words in the post-test, approaching the "final lowering" of native recordings. In addition, more pairs of neighbouring stressed and unstressed syllables showed significant differences in mean fundamental frequency in the post-test (6) vs. pre-test (2), suggesting an improved use of pitch to enhance the contrast of stressed and unstressed syllables.

Overall, this data indicates that participants showed improvements in some elements of English rhythm and intonation after the 10-week training. It should be noted that it is difficult to untangle the effects of shadowing and oral reading, as both were used as part of the intervention. However, as the steps of the procedure described could fit with what Hamada and Suzuki (2022) describe as text-presented shadowing, which is often used in combination with other forms of shadowing in pedagogical interventions (e.g., Hamada, 2021), this is not as problematic as it initially appears.

Sumiyoshi (2014) also reported positive results in two studies exploring the impact of shadowing on pitch accent amongst learners of Japanese. One study was a quasi-experimental study comparing improvements in pitch accent between a shadowing and non-shadowing group. Participants were 46 students of Japanese at an Australian university, enrolled in either Intermediate Spoken Japanese or Intermediate Japanese I. There were 26 females and 20 males, but no information is given about participants' age, L1, or majors. Participants were divided into a control (n = 34) and experimental group (n = 12). Due to university policy on assessment, the experimental group included students enrolled in both Intermediate Japanese I and Intermediate Spoken Japanese, and the control group included students enrolled only in Intermediate Japanese I. This resulted in a number of other differences between the control and experimental groups, such as overall numbers, gender balance, and teaching input hours. From the limited description available, it appears that the students in the experimental group received an additional 2 hours of instruction per week, and 7 weeks of shadowing homework over one semester. It is unclear what the control group did in their classes.

Recordings from shadowing homework in weeks 2 and 10 of the semester were analysed for pitch-accent accuracy. To address issues of bias and objectivity, participants IDs were replaced with numerical IDs and random numbers were given to the weeks. An online prosodic reading system was used to detect pitch accent falls in the model audio and provide a visual representation of model prosody. Sumiyoshi (2014) then appears to have rated students' recording against this model, although the exact method followed is unclear. Results indicated that mean scores in pitch accuracy for both the experimental and control groups improved in pitch accuracy from week 2 to week 10. Median scores in the experimental group and the control group were significantly higher at post-test than pre-tests (p=0.041 and p=0.295, respectively). However, the effect size for the experimental group was medium (d=0.64) and the effect size for the control group (d=0.15) was small. Whilst these

results do suggest that shadowing improved pitch accent, they should be taken with caution due to the significant differences between the two groups, and the use of a single rater.

Despite the issues in the abovementioned study, results do concur with those of several other articles reviewed. For example, in Sumiyoshi's (2014) other study, 20 students, aged 18-30 and only enrolled in spoken Japanese, also showed significant improvements in median scores of pitch accent (p=0.010, d=0.57) after 5 weeks of shadowing homework. Rongna and Hayashi's (2012) study also reported significant improvements (p<0.001) in pitch accent amongst 15 Chinese and Mongolian intermediate level learners of Japanese after 7 weeks of shadowing training. It should be noted, however, that both studies relied on a single rater, and that the ecological validity of Rongna and Hayashi's (2012) study is limited due to use of the same text during training and testing.

Kunihara et al.'s (2022) study, on the other hand, reported mixed results in terms of improvement in prosody. During a 40 day "shadowing marathon", participants showed significant improvement in pitch, duration, and intensity from day one to day 23 (p<.05), but not from day 20 to 42. However, this improvement was only observed in standard shadowing, not script shadowing or reading aloud.

4.2.2.2.3. Impressionistic and acoustic measurements

Only one study (Hori, 2008) used a combination of impressionist ratings and acoustic measurements to assess improvements in prosody. Unlike the abovementioned studies, results were mixed. Participants were 43 Japanese low-intermediate learners of English, divided into an experimental (n = 26) and control group (n = 17). The experimental group received 1 month of shadowing training, in which they shadowed twice a week, with each session lasting no more than one hour. Read-aloud pre-tests, post-tests, and delayed post-tests (1 month after training) were administered, and recordings were subjected to acoustic analysis. Results indicated no significant interactions between time and group for durational differences between stressed and unstressed vowels. For fundamental frequency (F0) ratios, only two out of five analysed words showed significant interactions between time and group, in favour of the experimental group (p<.05). This was similar for fundamental frequency range, in which only three out of five tone groups analysed showed significant interactions for time and group, in favour of the experimental group (p<.05). In addition, recordings were

rated impressionistically for prosody by two native speakers of English, but no significant interaction was observed between time and group (p=.139). Whilst the design of this study is highly appropriate, trustworthiness of results is limited by the fact that little information is provided about the intervention received by the control group, and whether groups were taught by the same teacher.

4.2.2.2.4. Unspecified measurements

Finally, two studies used unspecified measurements of prosodic improvement, and are therefore highly limited in trustworthiness. Hsieh et al. (2023) used a computer program, My English Teacher, of an unspecified nature, and reported that, amongst 14 English learners at Taiwan National University, and experimental shadowing group (n = 7) outperformed a control group (n = 7) on measures of intonation. This study also did not report the content of shadowing and control training. Nguyen and Nguyen (2019) also reported improvements in stress and intonation amongst 40 Vietnamese young learners of English, of an elementary proficiency level, after 8 weeks of shadowing. However, they failed to provide information on tasks used to collect speech samples or scoring measurements and procedures.

4.2.3. Improvements in segmental pronunciation control

Four studies focused on the effect of shadowing on development of segmental pronunciation control. The weight of evidence for these studies is assessed below:

Table 23. *Key information and weight of evidence segmental studies*

Study	Task type	Scoring method	Relevance RQ1	Appropriateness RQ1	Contribution RQ1	Trustworthiness RQ1
Haufe (2013)	Controlled	Human rater	High	Low	Low	Low
Huang (2018)	Controlled	Acoustic analysis	High	Medium	Medium	Medium
Kunihara et al. (2022)	Controlled	Acoustic analysis	High	Medium	Medium	Medium
Nguyen & Nguyen (2019)	Unknown	Unknown	High	Low	Low	Low

All studies used controlled tasks to collect pronunciation samples, limiting their appropriateness. Overall, results on the impact of shadowing on segmental control were mixed, with some studies reporting positive results and others reporting inconclusive results.

Two studies reported positive results. Haufe (2013) conducted a case study with one Chinese English learner studying at a university in Canada. The participant shadowed a recording of a class presentation for two weeks. From pre-test (a diagnostic test and interview) to post-test (a rehearsal and in-class performance of a presentation), the participant showed improvement in her inaccuracy rate of the /θ/ phoneme (from 100% at pre-test, to 9% in her rehearsal and 50% in her presentation) the /s/ phoneme in word-final position (from 33% at pre-test, to 20.9% in her rehearsal and 25% in her presentation). However, the pre-test only contained five instances of /θ/, all of which were in the word "think". It therefore cannot be considered an accurate measure of the participant's ability to deal with the phoneme in different words and word positions. This is particularly pertinent as the author notes that the participant mispronounced "thinking" at post-test, indicating it may be a particularly challenging word for her. In addition, using inaccuracy rates for word-final /s/ to measure pronunciation improvements in Chinese learners is complicated, given that problems with word-final /s/ amongst speakers of this language are of both a grammatical and phonetic nature (Yakovleva, 2013). Overall, given the methodological issues noted and the small sample size, these results are low in their trustworthiness.

The second study reporting improvements in segmental production was also low in trustworthiness, largely due to missing information. Nguyen and Nguyen (2019) conducted a study involving an 8-week shadowing intervention with 40 young Vietnamese learners of English, of an elementary proficiency level. Comparison from a pre- and post-test of pronunciation indicated that 72.5% of students improved in their pronunciation of exponents and consonants, 70% in vowels, and 67.5% in phonetic variants. However, no information is provided on the design, validation, components or scoring of the test, meaning that these results should be taken with great caution.

A further two studies reported more inconclusive results. For example, Huang's (2018) study of a 4-week shadowing programme amongst 20 Taiwanese beginner learners of English. Participants were aged 8-12 and spoke Taiwanese Mandarin as an L1. Participants were randomly divided into a control group (n = 10) and a treatment group (n = 10). However,

final composition of age and gender of the groups was unbalanced, with more older students and females in the experimental group. Participants from both groups continued to attend their regular English classes, and the treatment group received 5-8 minutes of shadowing training 4 days per week for 4 weeks.

In pre-tests and post-tests, participants read passages out loud and then orally answered questions about them. Answers involved orally filling one-word gaps in sentences. Both tests followed the same format, but the post-test was of a higher level of English than the pre-test, which was observed to be too easy for students. Test recordings were analysed for improvement in Voice Time Onset (VOT) for the consonants /b/, /p/, /d/, /t/ and /k/. No significant differences were observed between groups at pre-test, except that the control group pronounced /t/ significantly closer to standard values in the speaking condition. At post-test, in the reading condition, the experimental group pronounced /p/ with a significantly more native-like value than the control group, with a large effect size (r=0.82). There were no other significant differences between the groups in this condition. In the speaking condition, the experimental group showed significantly more improvement in their production of /b/, /p/ and /d/ than the control group. However, only the effect size value of /p/ (0.98) reached the recommended level of 0.8. In addition, despite the experimental group's improvement, the control group produced /b/ and /d/ as closer to standard averages. Overall, these results therefore provide an inconclusive picture of whether shadowing can improve VOT in production of consonants. Furthermore, the age and gender differences between the two groups, and the lack of counter-balanced tests, limit the trustworthiness of results.

Finally, Kunihara et al. (2022) also reported mixed results in a 42 day "shadowing marathon" amongst 20 Japanese university-level learners of English. Students shadowed four texts daily, all of which were selected from the EIKEN listening test to ensure similar difficulty. They performed three rounds of standard shadowing, one round of script shadowing, and then read the passage aloud. Acoustic analysis of students' recordings was used to determine segmental gaps between learner productions and model recordings. Within sessions, that is, between rounds of shadowing, learners' segmental control generally became significantly closer to the model control (p<.01). However, segmental control for script-shadowing and first readings aloud were significantly closer to the model (p<.001), indicating that articulatory control relied heavily on written text. Comparing the first round of standard shadowing to the model on day 1 and day 23, and day 20 with day 42, segmental control

became significantly closer to the model (p<.001). However, when comparing script-shadowing and first reading on day 1 and 21, and day 20 and 40, no significant differences were observed, indicating that shadowing was not effective for training pronunciation when text cues were available. It should be noted that no control group was included in the study, which could have helped unpack these inconclusive results further.

4.3. How do learners evaluate shadowing as a pronunciation teaching and learning technique?

4.3.1. Relevant studies

A total of 16 studies explored student perceptions of shadowing as a pronunciation training activity, and a further three studies explored student perceptions of different types of shadowing applications or software. However, these latter studies are not discussed due to space constraints and the fact that their results were considered fragmented and less relevant to the central research question. Critical reviews can be consulted in Annex 3.

The weight of evidence for the 16 studies was assessed as is shown in Table 24.

Table 24. *Key information and weight of evidence for perception studies*

Study	Data collection instruments	Relevance RQ2	Appropriateness RQ2	Contribution RQ2	Trustworthiness RQ2
Bovee and Stewart (2008)	Questionnaire	High	Medium	High	Medium
Foote and McDonough (2014)	Structured interviews	High	Medium	High	Medium
Hamada (2018b)	Semantic differentiation task	High	Medium	High	Medium
Haufe (2013)	Questionnaire, semi-structured interview, teacher observations	High	High	Medium	Medium

Horiyama (2012)	Questionnaire	High	Medium	High	Medium
Huang (2018)	Questionnaire	High	Medium	High	Medium
Huang et al. (2023)	Questionnaire, semi-structured interviews	High	High	High	Medium
Kuo and Chou (2014)	Questionnaire, teacher observations	High	High	High	Medium
Lin (2009)	Questionnaire, semi-structured interviews, teacher observations	High	High	High	Medium
Omar and Umehara (2010)	Questionnaire, student journals, participant observation	High	High	Medium	Medium
Saito et al. (2011)	Questionnaire	High	Medium	High	Medium
Sumiyoshi (2014)	Questionnaire	High	Medium	High	Medium
Mishima and Cheng (2017)	Questionnaire	High	Medium	Medium	Medium
Teeter (2017)	Questionnaire	High	Medium	High	Medium
Wang (2018)	Questionnaire	High	Medium	High	Medium
Willardson (2014)	Questionnaire, teacher observations	High	High	High	Medium

All studies were assessed as highly relevant to RQ2, due to a clear focus on student perceptions of shadowing in their research questions and study design. In addition, many studies were assessed as making a high contribution to the RQ2, as results, in general, provided informative data on learners' evaluations of shadowing. The contribution of studies was more limited in cases with small sample sizes (Haufe, 2013), opaque reporting (Omar & Umehara, 2010), or lack of clearly defined shadowing procedures (Mishima & Cheng, 2017).

However, it should be noted that fewer studies were assessed as highly appropriate. Those that did meet this criterion combined elements of data triangulation to collect data on perceptions, combining either surveys and interviews (Huang et al. 2023), surveys/interviews and classroom observation (Kuo & Chou, 2014; Lin, 2009; Willardson, 2014), or journals, questionnaires and participant observation (Omar & Umehara, 2010). They also all employed carefully defined shadowing procedures. Other studies tended to explore perceptions through surveys alone. As all studies were considered limited in trustworthiness for similar reasons, these limitations are explored collectively at the end of this section.

Results are organised below by the perceived interest and enjoyment of shadowing tasks, the perceived effectiveness of the technique, and how student perceptions of the technique changed over time.

4.3.2. Perceived interest and enjoyment in shadowing tasks

In general, studies reported an interest in shadowing amongst students. For example, in Wang's (2018) survey, with 40 Chinese university students of English of an unspecified proficiency level, 90% of students found the 10-week shadowing intervention "interesting". A similar finding was reported in Huang's (2018) survey of 10 Taiwanese young learners of elementary-level English, with 80% of students stating that shadowing was "very interesting" or "interesting" after 4 weeks of training. Many other studies reported an overall interest in and enjoyment of shadowing (Bovee & Stewart, 2008; Foote & McDonough, 2017; Kuo & Chou, 2014; Lin, 2009; Omar & Umehara, 2010; Willardson, 2014), with one study suggesting that novel forms of shadowing, like haptic shadowing, in which participants "punch" stressed syllables, may be perceived particularly positively (Hamada, 2018b). It should be noted that positive experiences of the technique were reported across a range of age groups, from adults (Bovee & Stewart, 2008; Foote & McDonough, 2014; Hamada, 2018b; Omar & Umehara, 2010) to teens (Lin, 2009; Willardson, 2014) and children (Kuo & Chou, 2014).

However, two studies reported less student interest in the technique. For example, Teeter's (2017) survey of 1001 first-year intermediate-level students of English at a university in Japan reported student boredom when shadowing. The survey involved 6-point Likert scale

questions, exploring students' enjoyment of shadowing after 12 weeks of shadowing homework. The median score for the item "shadowing is interesting" was 3 (close to slightly disagree), whilst median scores for the items "shadowing is boring" and "shadowing is an effective way of learning English" were 4 (close to slightly agree). It should be noted, however, that reliability analyses indicated that between 8.76% and 29.09% of respondents may not have been completing the survey in a consistent fashion with respect to these two items. In a similar vein, in Saito et al.'s (2011) survey of 41 mixed-proficiency third year high school students of English in Japan, only 38% of students stated they found the three-class intervention fun.

The concerns shared by students in the two abovementioned studies were also echoed by a minority of students in studies reporting overall enjoyment of shadowing. The most frequently mentioned issues with the technique were frustration with the speed of audio materials (Horiyama, 2012; Huang, 2018; Sumiyoshi, 2014) and the time-consuming, repetitive nature of tasks (Huang et al., 2023; Lin, 2009). Some participants also complained about feeling bored by shadowing (Huang, 2018; Lin, 2009), procrastination surrounding the task (Sumiyoshi, 2014), or feeling demotivated because of failure (Lin, 2009). More specific difficulties, such as shadowing unfamiliar vocabulary and connected speech, were also reported (Omar & Umehara, 2010).

One survey (Sumiyoshi, 2014), conducted with 36 advanced learners of Japanese studying at an Australian university, also indicated that complaints surrounding shadowing can be more frequent amongst lower proficiency students. These students gave more negative comments in open response sections than their higher proficiency counterparts (22 vs. 18, respectively), and commented more on the difficult speed of the audio (12 vs. 6 comments).

4.3.3. Perceived effectiveness of shadowing

As a whole, studies reported that students perceived shadowing as an effective method to train different language skills.

Some studies focused on the value of shadowing for oral English skills and improving speaking proficiency. For example, in Sumiyoshi's (2014) survey, the details of which are presented in section 4.3.2, 80% of participants considered shadowing effective for developing

speaking skills. In Horiyama's (2012) survey of 25 university students learning intermediate-level English in Japan, this figure was even higher, at 90%.

Other studies highlighted the effectiveness of shadowing for training the sub-skill of pronunciation. For instance, in Kuo and Chou's (2014) survey of 53 fourth-grade Taiwanese English learners, of an unspecified proficiency level, 82.7% of students believed that a 12-week shadowing was a useful way to improve pronunciation. This was echoed in Haufe's (2013) case study of one Chinese ESL student at a Canadian university, with the participant rating the activity highly (7/10, with 10 being "very high") as a technique to improve her pronunciation. Finally, in Huang et al.'s (2023) survey of 90 Taiwanese university learners of English, with proficiency ranging from A2-B1, students strongly agreed that 6 weeks of shadowing practice was useful to improve both speaking and pronunciation.

A range of reasons were given for the perceived effectiveness of shadowing for training these skills. In Haufe's case study, the participant highlighted in an interview that it allowed her to compare her pronunciation with a model and notice her mistakes. In Huang et al.'s (2023) study, semi-structured interviews with 18 randomly selected participants indicated that students valued the intervention as it raised their awareness of pronunciation and helped them improve word stress, intonation, and fluency. Horiyama's (2012) survey also indicated that participants viewed shadowing as effective or highly effective for training prosody (80%). Similarly, in Sumiyoshi's (2014) survey of 20 intermediate learners of Japanese at an Australian university, after a 7-week shadowing homework programme, participants believed shadowing was effective for developing fluency and speed (80%), pitch accent (78.3%) and native-like speech (61.7%). Participants in a survey of nine non-native speakers of English at a US university also appeared to view their 2-week shadowing programme as effective for improving specific sub-skills like fluency, rhythm and intonation (Mishima and Cheng, 2017). However, as shadowing in this last study was combined with use of a novel programme GoAnimate, in which students created avatars to give final presentations, it is difficult to disentangle student perceptions of each component of the intervention.

4.3.4. Reported changes in perceptions over time

Four studies reported a change in attitudes to shadowing over time, with student views becoming more positive the more they shadowed. Lin's (2009) data, from classroom observations and semi-structured interviews with 25 Taiwanese junior high students of elementary-level English, reported initial difficulties and fears of shadowing which were gradually overcome over the course of a 5-week programme. Omar and Umehara's (2010) 6-month action-research project highlighted similar findings amongst four retired Japanese learners of English living in Malaysia. The authors report that, in diaries, participants initially showed mixed feelings about shadowing, but became more enthusiastic and interested by the end of the study. However, it should be noted that no excerpts are provided to evidence this claim. Increasingly positive attitudes to shadowing were also reported in two surveys: one of 18 American high school students of intermediate-level French, after a 10-week shadowing programme (Willardson, 2014), and the other of 16 advanced ESL learners in Canada after an 8-week programme (Foote & McDonough, 2017). It should be highlighted that there may be an element of self-selection bias in this last study, as six participants left the study, potentially because they may not have perceived shadowing positively.

Similarly, in several studies participants expressed an interest in continuing shadowing after the intervention had finished. 64%, 73.1% and 85% of participants, in Sumiyoshi (2014), Wang (2018) and Kuo and Chou's (2014) surveys, respectively, expressed an intention to continue to use shadowing.

4.3.5. Limitations of perceptions research

The body of research presented above has several significant limitations, as is evident in the ratings for appropriateness and trustworthiness of many of the studies (see Table 24).

The first limitation is related to data collection instruments. The majority of the studies (10/16) exploring learner evaluations of shadowing relied purely or heavily on questionnaires or semantic differentiation tasks, which are limited in their ability to provide rich data on student experiences and perceptions. Only five studies used interviews, with four studies using semi-structured interviews (Haufe, 2013; Huang et al., 2023; Lin, 2009; Mishima & Cheng, 2017) and one employing structured interviews (Foote & McDonough, 2017). In

addition, as mentioned, only four studies combined different instruments for data triangulation on perceptions.

Secondly, in the bulk of survey-based studies, the validity and reliability of questionnaires was not addressed. The only studies involving discussions of questionnaire validity included Lin (2009), whose questionnaire was validated by an expert professor and two teachers, and Huang et al. (2023), whose questions were checked by four expert instructors. Only three studies included measurements of consistency or reliability, with Sumiyoshi (2014) and Huang et al.'s (2023) questionnaires showing high internal consistency as measured by Cronbach's alpha (.858 and .84, respectively). Hamada's (2018) study also addressed the issue of reliability through use of a questionnaire already tested for reliability in previous work (Hamada, 2011). Use of a pre-existing tested questionnaire also allowed for greater comparability with other research, something lacking in other studies due to their use of self-designed questionnaires or items on shadowing.

In terms of data analysis, when qualitative instruments, like semi-structured interviews, diaries, or classroom observation, were employed, information on methods of data analysis was sometimes not reported (in the cases of Haufe, 2013; Huang et al., 2023), or underdeveloped (in the case of Foote & McDonough, 2014). Only Lin (2009), Mishima and Cheng (2017), and Omar and Umehara (2010) gave full explanations of the coding processes used to analyse their qualitative data.

Finally, it should be noted that many studies may show some inflation of positive perceptions due to the fact that survey responses may not have been anonymous, and the fact that data collection may have been carried out by teachers and teacher-researchers. Only two survey-based studies mention that students were informed their responses would be anonymous (Hamada, 2018; Mishima & Cheng, 2017). Lack of guaranteed anonymity in other studies could have led students to present overly positive views of interventions to please teachers and researchers and avoid perceived negative repercussions of critiquing superiors. In a similar vein, most studies made no explicit mention of who administered questionnaires or conducted interviews. In those that did, data was collected by teachers or teacher-researchers (Haufe, 2013; Hamada, 2018; Lin, 2009). As most other surveys and interviews appear to have been conducted in or after class, with no mention of employing additional researchers, it is likely that they were also administered by teachers or teacher-

researchers. If this is indeed the case, students may have been hesitant to express more negative views to those who had taught them shadowing.

5. Discussion and conclusion

This section provides an overview of the results, per research question, and highlights key methodological and theoretical limitations of the body of research reviewed. It uses these limitations to provide suggestions for future research into shadowing for pronunciation. The limitations of the review itself are also presented.

5.1. RQ1 Shadowing and pronunciation improvement

5.1.1. Strength of evidence per pronunciation feature and areas for improved research

Overall, the research reviewed tentatively suggests that shadowing can contribute to improving L2 learners' pronunciation proficiency. However, the quality and conclusiveness of evidence varied per feature of pronunciation studied, and was related to a number of methodological issues.

Research into comprehensibility, intelligibility, and accentedness was relatively well developed, with seven out of nine studies reporting positive results and two studies considered highly appropriate, relevant, and able to contribute to RQ1 (Foote & McDonough, 2017; Shao et al., 2023). However, these positive results are limited by the fact that the majority of studies (7/9) used only controlled speaking tasks, which may provide an inflated picture of overall pronunciation development. An overreliance on such controlled tasks has been observed in pronunciation research as a whole, and has been noted as problematic due to concerns about ecological validity (Lee et al., 2015; Thomson & Derwing, 2015). That is, improved performance in controlled tasks, such as read-aloud tests, may not directly translate into meaningful improvement in real-world speaking, in which learners are not able to focus solely on pronunciation control and must attend to communication and retrieval of vocabulary and grammar (Thomson & Derwing, 2015). Indeed, research suggests that pronunciation production tends to be more target-like in formal, controlled tasks, compared to spontaneous ones (Major, 2008; Rau et al., 2009; Saito & Brajot, 2013).

As well as prompting different levels of pronunciation control, controlled and spontaneous tasks also tap into different types of pronunciation knowledge. Whilst controlled tasks elicit more conscious application and monitoring of declarative knowledge, spontaneous tasks

elicit more automatised and proceduralised knowledge (Saito & Plonsky, 2019). Consequently, recent research argues that controlled and spontaneous L2 pronunciation performance are two distinct phenomena and should therefore be assessed and interpreted separately (Saito & Plonsky, 2019). To provide a more valid, representative picture of learners' pronunciation improvements, future studies into shadowing should therefore include both controlled and spontaneous measures, as has been recommended for pronunciation research as a whole (Thomson & Derwing, 2015).

It should also be noted that there was some conceptual "fuzziness" in a number of intelligibility and comprehensibility studies. One study (Mishima & Cheng, 2017) referred to intelligibility, the degree to which a speaker's message is understood, when what was actually measured was comprehensibility, or ease of listener understanding. In addition, two other studies (Willardson, 2014; Hori, 2008) merged two or more constructs in the same scale, making it difficult to disentangle the effect of shadowing on each independent construct. Future studies on shadowing should therefore aim to measure constructs separately and respond to calls to operationalise measurements in line with original definitions (Thomson, 2017). Furthermore, as intelligibility was under researched in the sample, more work exploring the impact of shadowing on this feature would be beneficial, using recommended measurements like listener transcriptions or cloze exercises (Thomson, 2017).

A substantial number of studies also explored the impact of shadowing on suprasegmental pronunciation control, both in terms of fluency and various aspects of prosody. All eight studies related to fluency reported positive results. Although more mixed, the 11 studies exploring prosody were also generally positive, reporting improvements in intonation and rhythm after shadowing (Kuo & Chou, 2014; Omar & Umehara, 2010; Ono et al., 2012), production of weak forms of function words (El-Esery, 2021; Nakayama & Armstrong, 2011) and use of pitch in English (Mori, 2011; Hori, 2008) and Japanese (Sumiyoshi, 2014; Rongna & Hayashi, 2012). The focus on suprasegmentals in the sample is noteworthy, as it is an area of pronunciation that is often under researched (Thomson & Derwing, 2015). Furthermore, suprasegmentals have been noted as important in listener judgments of comprehensibility and accentedness (Kang, 2010), and are therefore important to help students develop pronunciation proficiency.

However, results related to suprasegmental control are also limited by several notable methodological features. Firstly, like the holistic studies, they showed an over reliance on controlled tasks, limiting ecological validity as explained above. In addition, there was a similarly problematic over reliance on acoustic measurements, with few studies (Hori, 2008; Rongna et al., 2013, 2015) exploring how measurable acoustic changes could be extrapolated to real listener judgments or measures of global pronunciation improvement. The question therefore remains as to the extent to which documented suprasegmental improvements would be perceptible to listeners and therefore constitute a significant improvement to participants' pronunciation. To address this issue, future research should aim to combine acoustic measurements with impressionistic judgments, as has been recommended in pronunciation research overall (Kang & Kermand, 2017; Thomson & Derwing, 2015). Future research into fluency may also benefit by triangulating existing data, which relied on measures of speech rate, with other valid measures of fluency, such as mean length of runs and phonation-time ratio (Kormos, 2006; Kormos & Dénes, 2004).

Finally, research into improvements in segmental pronunciation control was the most limited area of study, both in terms of the number of studies available and in terms of methodological issues. Only four studies explored segmentals, with two reporting positive results and two reporting mixed results. In addition, both studies reporting positive results were considered low in appropriateness, trustworthiness, and contribution, largely due to undescribed or invalid pre- and post-tests. Overall, the evidence for the impact of shadowing on segmentals is therefore inconclusive and should be developed in future research to provide a fuller picture of the impact of the technique.

Whilst, as a whole, results do indicate that shadowing can improve various aspects of L2 pronunciation control, it should be noted that the studies reviewed tended to focus on university level learners of English. More research into learners of other L2s, at different educational levels, is therefore needed to increase the generalisability of results. It would also be beneficial to explore the impact of shadowing on features of connected speech (for example elision, or linking), which was not reviewed in a meaningful way in any of the studies reviewed and is an area in need of development in pronunciation research (Lee et al., 2015). In addition to these research gaps, several additional methodological and theoretical limitations, recurrent across all the features of pronunciation control explored, temper the overall findings. These issues are discussed below.

5.1.2. Additional methodological limitations

Firstly, several studies in the sample used the same materials for pre-test, post-test, and training (e.g. Rongna et al., 2013; 2015). This is not advisable, as it provides no information on whether learners can generalise what they have learnt (Thomson & Derwing, 2015). Other studies did not report the content of one or both tests (Hsieh et al., 2023; Nguyen & Nguyen, 2019; Wang, 2018) or used identical pre- and post-tests with no consideration of how this may have affected results (e.g. Foote & McDonough, 2017; Mori, 2011; Willardson, 2014; Yavari & Shafiee, 2018). Ideally, tests should be different, to avoid repetition effects, but counterbalanced to control for the variables of topic or language difficulty. Shao et al.'s (2023) study is a good example to follow in this regard.

Secondly, whilst many studies used multiple raters, reporting their L1, background and calculations of interrater reliability, few studies provided information on how bias was reduced in the rating process. In particular, few studies (e.g. Foote and McDonough, 2017; Sumiyoshi, 2014; Willardson, 2014) clearly specified whether raters were blind with regard to group or time point of recordings. Blind rating procedures should be implemented to increase internal validity in future research.

Finally, it should be noted that only one study (Hori, 2008) included a delayed post-test. This has been reported as an issue in L2 pronunciation research in general (Thomson & Derwing, 2015), and is problematic as pronunciation interventions should ideally be able to provide long-term results. It is therefore crucial to include more delayed post-tests in further research, to assess the extent to which shadowing can provide durable improvements in pronunciation.

The issues noted above are in line with similar observations from leading scholars in the field of shadowing. Hamada (2019), for example, has highlighted the lack of robust research design as a key problem facing research on shadowing and speaking.

5.1.3. Additional theoretical limitations

The body of research reviewed also presented some more theoretical limitations. One key point to note is that a number of studies (Althubyani, 2021; Hashimoto et al., 2022; Hutchinson, 2022; Nakayama, 2021; Rojczyk, 2013, Šturm et al., 2022, Zając & Rojczyk, 2014, Zhang & Peng, 2017) conflated shadowing with repetition. As outlined in Chapter 2, the two techniques are actually significantly different and, due to differences in their online and offline nature, respectively, are thought to work in different ways. Future studies aiming to explore shadowing should be highly sensitive to this issue, and design interventions in accordance with established definitions of shadowing. Shadowing procedures should also be carefully reported, to allow for evaluation of the degree to which they respect or deviate from current definitions. Such information was missing in a number of studies (e.g., Kurniawan et al., 2019; Mishima and Cheng, 2017; El-Esery, 2021).

In addition, of studies with enough data to determine their alignment with Nativeness or Intelligibility Principles (Levis, 2005, 2020), more studies were implicitly in line with the former (12) than the latter (2). Four studies showed elements of both principles. The predominance of studies in line with the Nativeness Principle can be considered problematic, as there is a consensus in the literature that L2 pronunciation instruction (Levis, 2005, 2020) and assessment (Harding, 2017; Kang & Kermad, 2017) should be guided by measures of comprehensibility and intelligibility. This is particularly important given that the majority of studies in this review involved learners of English, a pluricentric language often used in lingua franca contexts. The appropriateness of comparison with only native or native-like pronunciation targets for current or future English as a Lingua Franca (ELF) users is questionable (Harding, 2017; Kang & Kermad, 2017).

For these reasons, in future studies on shadowing for pronunciation, researchers should reflect more consciously on their theoretical orientation, and strive to include more measures of comprehensibility and intelligibility. Several studies from this review (Foote and McDonough, 2017; Huang et al., 2023; Shao et al., 2023) are examples to follow in this respect. In addition, when discrete measures of segmental and suprasegmental control are used, there should be more theorisation and explanation of how these features lead to global improvement in comprehensibility and intelligibility (Saito & Plonksy, 2019). In the case of English, comprehensibility research highlighting principles of functional load (Catford, 1987;

Jenkins, 2000), and the importance of nuclear stress (Hahn, 2004) and word stress (Field, 2005) could also be used to guide the selection of target segmental and suprasegmental features in shadowing interventions. More use of non-native raters, following the examples of several studies in the review (e.g. Hamada, 2018; Shao et al., 2023), would also be valuable in studies in which participants are English learners who will ultimately go on to use the language in ELF situations.

5.2. RQ2 Shadowing and student perceptions: areas for improved research and practice

In general, this review also indicates that learners view shadowing favourably, with the majority of the 16 studies exploring perceptions reporting mostly positive attitudes towards the technique. Many learners appeared to find the technique an interesting, enjoyable, and valuable way to improve general speaking and pronunciation skills, as well as sub-skills like fluency, word stress, and elements of prosody. These results suggest that the technique may be a welcome addition to pronunciation training in the classroom.

However, results are tempered by the fact that most surveys and interviews are likely to have been carried out by teachers, which could have inflated positive perceptions shared by students. The lack of research exploring perceptions through rich, qualitative data, and data triangulation combining different data collection instruments (e.g. a combination of classroom observation, interviews, journals) is also noteworthy. Future research into perceptions of shadowing should ideally involve such data triangulation, as well as data collection by researchers not involved in teaching and with no position of power compared to learners.

Despite overall positive evaluations of the technique, several negative aspects of shadowing did emerge from the data, for example difficulties with the speed of audio, frustration with the time-consuming, repetitive nature of training, and boredom. These issued are echoed in research into shadowing for listening (Hamada, 2017). Teachers and researchers should, therefore, be aware of these challenges and take steps to lessen their impact on students. For instance, starting with slow audio speeds, and potentially providing different speeds for different proficiency levels (Sumiyoshi, 2014), as well as selecting appropriate speech models for learners (Foote, 2017). To tackle boredom amongst students, shadowing

interventions could allow for learner autonomy in selection of materials, and provide different types of materials for different age groups. For example, using chants and songs for young learners (Kuo and Chou, 2014) or self-selected cultural materials for adolescent learners (Willardson, 2014).

Finally, it should be noted that the majority of studies involving perceptions research were carried out with Chinese (Haufe, 2013; Wang, 2018), Taiwanese (Huang, 2018; Huang et al., 2023; Kuo & Chou, 2014; Lin, 2009), and Japanese (Bovee & Stewart, 2008; Hamada, 2018; Horiyama, 2012; Omar & Umehara, 2010; Saito et al., 2011; Teeter, 2017) learners of English. This is a promising start to filling the research gap of student perceptions of the technique outside Japan (Hamada, 2018). However, more research is needed on perceptions of the technique in other national contexts and "cultures of learning" (Cortazzi & Jin, 1996; 2013), as well as in other "small cultures" in different types of communities, organisations, and social settings (Holliday, 1999). There is also a need for more research amongst learners of more diverse L1s and L2s. Examples like Willard's (2014) study of American learners of French, Sumiyoshi's (2014) study on mixed L1 learners of Japanese in Australia, or Foote and McDonough's (2017) study of mixed L1 ESL learners in Canada, are examples to follow in these regards.

5.3. Limitations of this review

There are several limitations to the present review. Firstly, in terms of scope, limiting the search to English language papers may have excluded some significant studies published in other languages. In backward citation checking, for example, a number of papers published in Japanese were encountered. Given that shadowing research started in Japan, and that the country is still a stronghold of shadowing research, future reviews should aim to include articles written in Japanese. Future reviews should also consider how to access research not covered by the databases used. Despite steps taken to include a broad of databases, including those containing grey literature, a preliminary search on Google Scholar revealed a significant number of articles related to shadowing that did not appear in the selected databases.

Secondly, given that this review is an MSc dissertation project, it was only possible to use a second reviewer for 10% of title and abstract screening. Ideally, the whole process of

screening and data extraction would have been conducted by both reviewers to minimise bias.

Finally, the focus of this review was limited to whether shadowing improves pronunciation, and how students perceive the technique. The review did not systematically explore the questions of *how* and *why* the technique is beneficial for pronunciation development, which are equally important. More research into this area, and development of a theoretical framework to explain speaking development through shadowing, is therefore needed (Hamada, 2019a).

5.4. Conclusion

Overall, this systematic review indicates that, beyond its established contribution to developing L2 listening skills (Hamada, 2017; Kadota, 2019), shadowing can also help learners develop L2 pronunciation control. Results suggest that shadowing can help improve comprehensibility, intelligibility, accentedness, and fluency. There is also evidence to suggest, albeit more tentatively, that shadowing can help develop specific elements of prosodic control, such as rhythm, intonation, production of weak forms in English, and use of pitch. In addition to this, learners appear to generally view the technique as interesting, enjoyable, and effective.

However, there are a number of notable methodological issues that temper these findings, with more research involving both spontaneous and controlled speaking tasks, connections between acoustic measurements and listener judgments, delayed post-tests, and rich qualitative data, amongst other measures, needed to increase their robustness. More studies amongst learners of more diverse L1s, L2s, and educational and cultural settings is also essential to increase generalisability. Finally, there is also a need for more engagement with the current theoretical paradigms in pronunciation research, which foreground intelligibility and comprehensibility, rather than nativeness.

Despite these issues, results do suggest that shadowing can be a beneficial technique to help instructors with the challenge of integrating pronunciation work into teaching practice, and to help learners work towards more intelligible, comprehensible pronunciation.

References

Althubyani, R. (2021). *The effect of shadowing in learning L2 segments: a perspective from phonetic convergence.* [Unpublished master's thesis]. University of Wisconsin Milwaukee.

Atkinson, K., Koenka, A., Sanchez, C., Moshontz, H., & Cooper, H. (2014). Reporting standards for literature searchers and report inclusion criteria: making research syntheses more transparent and easy to replicate. *Research Synthesis Methods*, 6, 87-95.

Baker, A. (2011). *Pronunciation pedagogy: second language teacher cognition and practice*. [Unpublished doctoral dissertation]. Georgia State University.

Bovee, N., & Stewart, J. (2008). The utility of shadowing. In A. Stoke (Ed.), *JALT2008 Conference Proceedings* (pp. 888-900). JALT.

Breitkreutz, J., Derwing, T., & Rossiter, M. (2011). Pronunciation teaching practices in Canada. *TESL Canada Journal, 19,* 51-61.

Burgess, J., & Spencer, S. (2000). Phonology and pronunciation in integrated language teaching and teacher education. *System, 28*(2), 191-215.

Buss, L. (2013). Pronunciation from the perspective of pre-service EFL teachers: an analysis of internship reports. In J. Levis & K. LeVell (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference* (pp. 255-264). Iowa State University.

Catford, J. (1987). Phonetics and the teaching of pronunciation: a systematic description of English phonology. In J. Morley (Ed.), *Current perspectives on pronunciation: practised anchored in theory* (pp. 87-100). TESOL.

Celce-Murcia, M. Brinton, D., & Goodwin, J. (2006*). Teaching pronunciation: a reference for teachers of English to speakers of other languages.* Cambridge University Press. (Original work published 1996).

Couper, G. (2016). Teacher cognition of pronunciation teaching: teachers' concerns and issues. *TESOL Quarterly, 51*(4), 820-843.

Couper, G. (2019). Teachers' cognitions of corrective feedback on pronunciation: their beliefs, perceptions and practices. *System*, 84, 41-52.

Couper, G. (2020). Pronunciation teaching issues: answering teachers' questions. *RELC Journal, 52*(1), pp. 128-143.

Cortazzi, M., & Jin, L. (1996). Cultures of learning: language classrooms in China. In H. Coleman (Ed.), *Society and the language classroom* (pp. 169-206). Cambridge University Press.

Cortazzi, M., & Jin, L. (2013). Introduction: researching cultures of learning. In M. Cortazzi, & L. Jin (Eds.) *Researching cultures of learning* (pp. 1-17). Palgrave Macmillan.

Darcy, I., Rocca, B., & Hancock, Z. (2021). A window into the classroom: how teachers integrate pronunciation instruction. *RELC Journal, 52*(1), 110-127.

Derwing, T. (2003). What do ESL students say about their accents? *The Canadian Modern Language Review*, 59(4), 547-566.

Derwing, T. (2010). Utopian goals for pronunciation teaching. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 1st Pronunciation in Second Language Learning and Teaching Conference* (pp. 24-37). Iowa State University.

Derwing, T. (2019). Utopian goals for pronunciation teaching revisited. In J. Levis, C. Nagle & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching conference* (pp. 27-35). Iowa State University.

Derwing, T., & Munro, M. (2009). Putting accent in its place: rethinking obstacles to communication. *Language Teaching, 42*(4), 476-490.

Derwing, T., & Rossiter, M. (2002). ESL learners' perceptions of their pronunciation needs and strategies. *System,* 30, 155-166.

Derwing, T., Diepenbroek, L., & Foote, J. (2012). How well do general-skills ESL textbooks address pronunciation? *TESL Canada Journal, 30*(1), 22-44.

Derwing, T., Munro, M., Thomson, R., & Rossiter, M. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition, 31*(4), 533-557.

Edwards, J., Chan, K., Lam, Y., & Wang, Q. (2021). Social factors and the teaching of pronunciation: what the research tells us. *RELC Journal, 52*(1), 35-47.

Ekayati, R. (2020). Shadowing technique on students' listening word recognition. *Indonesian Journal of Education and Mathematical Science, 1*, 1-11.

El-Esery, A. (2021). English suprasegmentals for EFL learners: an experimental study with shadowing technique. *Asian EFL Journal Research Articles, 28*(3.2), 54-71.

English with Lucy. (2023, September 7). *English speaking method- speak with me! (Shadowing method)* [Video]. YouTube. https://www.youtube.com/watch?v=N20jOJDYyYA

English like a Native. (2018, April 27). Perfect English pronunciation: shadowing lesson 1. YouTube. https://www.youtube.com/watch?v=AeFzU-cv94s

Field, J. (2005). Intelligibility and the listener: the role of lexical stress. *TESOL Quarterly, 39*(3), 399-423.

Foote, J. (2017). *Shadowing: a useful pronunciation practice activity. What is shadowing?* Pronunciation for teachers. https://www.pronunciationforteachers.com/uploads/6/0/5/9/60596853/teaching_techniques_shadowing_jfoote.pdf

Foote, J., Holtby, A., & Derwing, T. (2011). 2010 survey of pronunciation teaching in adult ESL programs in Canada. *TESL Canada Journal, 29*(1), 1-22.

Foote, J., Trofimovich, P., Collins, L., & Soler Urzúa, F. (2016). Pronunciation teaching practices in communicative second language classes. *The Language Learning Journal, 44*(2), 181-196.

Foote, J., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation, 3*(1), 34-56.

Giustini, D., & Boulos, M. (2013). Google Scholar is not enough to be used alone for systematic reviews. *Online Journal of Public Health Informatics*, *5*(2), 214.

Gøtzche, P. (2022). Citation bias: questionable research practice or scientific misconduct? *Journal of the Royal Society of Medicine, 115*(1), 31-35.

Gough, D., Oliver, S., & Thomas, J. (2017). Introducing systematic reviews. In D. Gough, S. Oliver & J. Thomas (Eds.), *An introduction to systematic reviews* (2nd ed., pp. 1-18). Sage.

Guo, S., & Hsu, T. (2017). Integration of peer assessment and shadowing strategies for improving the oral performance of EFL learners. In W. Chen et al. (Eds.), *Proceedings of the 25th International Conference on Computers in Education* (pp. 928-230). Asia-Pacific Society for Computers in Education.

Hamada, Y. (2016a). Shadowing: who benefits and how? Uncovering a booming EFL technique for listening comprehension. *Language Teaching Research, 20*(1), 35-52.

Hamada, Y. (2016b). Wait! Is it really shadowing? *The Language Teacher, 40*(1), 14-17.

Hamada, Y. (2017). *Teaching EFL learners shadowing for listening: developing learners' bottom-up skills.* Routledge.

Hamada, Y. (2018a). Shadowing for language teaching. *CONTACT magazine*, 19-24.

Hamada, Y. (2018b). Shadowing for pronunciation development: haptic-shadowing and IPA-shadowing. *The Journal of Asia TEFL, 15*(1), 167-183

Hamada, Y. (2019a). Shadowing: what is it? How to use it. Where will it go? *RELC Journal, 50*(3), 386-393.

Hamada, Y. (2019b). Shadowing: for better understanding accented Englishes. *The Journal of Asia TEFL*, 16(3), 894-905.

Hamada, Y. (2021). Shadowing procedures in teaching and their future. *The Language Teacher Online, 45*(6), 32-36.

Hamada, Y. (2022). Developing a new shadowing procedure for Japanese EFL learners. *RELC Journal, 53*(3), 490-504.

Hamada, Y., & Suzuki, S. (2020). Listening to Global Englishes: script-assisted shadowing. *International Journal of Applied Linguistics, 31*(1), 31-47.

Hamada, Y., & Suzuki, Y. (2022). Situation shadowing in a framework of deliberate practice: a guide to using 16 techniques. *RELC Journal, 0*(0), 1-9.

Han, L. (2004). Primary stress and intelligibility: research to motivate the teaching of suprasegmentals. *TESOL Quarterly, 38*(2), 201-223.

Hashimoto, D., Gomi, K., & Shiraishi, R. (2022). Updating L2 phonetic memories: shadowing training vs. listening training. *Journal of the Phonetic Society of Japan, 26*, 13-26.

Harding, L. (2017). Validity in pronunciation assessment. In L. Harding (Ed.), *Assessment in second language pronunciation* (pp. 30-48). Routledge.

Haufe, H. (2013). *A case study of shadowing as a means of helping EAP students to prepare for oral presentations: effects on pronunciation and anxiety.* [Unpublished master's dissertation]. Carleton University.

Henderson, A., Frost, A., Tergujeff, E., Kautzsch, A., Murphy, D., Kirkova-Naskova, A., Waniek-Klimczak, E., Levey, D., Cunningham, U., & Curnick, L. (2012). The English

pronunciation teaching in Europe survey: selected results. *Research in Language, 10*(1), 5-27.

Hişmanoğlu, M. (2006). Current perspectives on pronunciation learning and teaching. *Journal of Language and Linguistic Studies, 2*(1), 100-110.

Holliday, A. (1999). Small cultures. *Applied Linguistics, 20*(2), 237-264.

Hori, T. (2008). *Exploring shadowing as a method of English pronunciation training.* [Unpublished master's dissertation]. Kwansei Gakuin University.

Horiyama, A. (2017). The development of English language skills through shadowing exercises. *Bunkyogakuindaigaku Gaikokugogakubu Bunkyogakuin-tankidaigaku Kiyo* [Bunkyo Gakuin Faculty of Foreign Studies Junior College Bulletin], 12, 113-123.

Hsieh, K., Dong, D., & Wang, Y. (2013). A preliminary study of applying shadowing technique to English intonation instruction. *Taiwan Journal of Linguistics*, 11.2, 43-66.

Hu, G. (2002). Potential cultural resistance to pedagogical imports: the case of communicative language teaching in China. *Language Culture and Curriculum, 15*(2), 93-105.

Huang, Y. (2018). *The influence of speech shadowing on English word-initial consonants produced by speakers of English as a foreign language.* [Unpublished master's dissertation]. California State University.

Huang, H., Barrett, N., Lo, M., & Tseng, C. (2023). The effectiveness of shadowing practice with web-based apps: towards promoting the comprehensibility and lexical-level intelligibility of EFL students' presentations. *English Teaching and Learning.* https://doi.org/10.1007/s42321-023-00145-w

Hurley, K. (2024, April 22). *An introduction to the shadowing technique*. FluentU. https://www.fluentu.com/blog/language-shadowing/#toc_1

Hutchinson, A. (2022). Individual variability and the effect of personality on non-native speech shadowing. *JASA Express Lett, 2*(6). https://doi.org/10.1121/10.0011753

Jones, R. (1997). Beyond “listen and repeat”: pronunciation teaching materials and theories of second language acquisition. *System, 25*(1), 103-112.

Kadota, S. (2019). *Shadowing as a practice in Second Language Acquisition: connecting inputs and outputs.* Routledge.

Kang, O., & Rubin, D. (2014). Listener expectations, reverse linguistic stereotyping, and individual background factors in social judgements and oral performance assessment. In J. Levis & A. Moyer (Eds.), *Social dynamics in second language accent* (pp. 239-253). De Gruyter Mouton.

Kang, O., & Kermad, A. (2017). Assessment in second language pronunciation. In O. Kang, R. Thomson, & J. Murphy (Eds.), *The Routledge handbook of contemporary English pronunciation* (pp. 511-526). Routledge.

Kato, S. (2009). Kokueigo noryoku shomei shutoku wo mezashita listening shido no kosatsu [Listening activities for the acquisition of Aviation English proficiency test]. *Bulletin of Chiba University Language and Culture,* 3, 47–59.

Kelly, G. (2013). *How to teach pronunciation.* Pearson. (Original work published 2000)

Kormos, J. (2006). *Speech production and second language acquisition*. Taylor & Francis Group.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145-164.

Kunihara, T., Zhu, C., Minematsu, N., & Nakanishi, N. (2022). Gradual improvements observed in learners’ perception and production of L2 sounds through continuing shadowing practice. *Interspeech 2022*, 1303-1307.

Kuo, Y., & Chou, T. (2014). Effects of text shadowing on Taiwanese EFL children’s pronunciation. *Asian EFL Journal, 16*(2), 11-43.

Kurniawan, H., Sitohang, B., & Rukmono, S. (2019). Gamification of mobile-based Japanese language shadowing. *Proceedings of the 2019 International Conference of Artificial Intelligence and Information Technology*, 215-219.

Lambert, S. (1992). Shadowing. *Meta, 37*(2), 263-273.

LaScotte, D., Meyers, C., & Tarone, E. (2021). Voice and mirroring in SLA: top-down pedagogy for L2 pronunciation instruction. *RELC Journal,* 51(2), 144-154.

LaScotte, D., & Tarone, E. (2022). Channelling voices to improve English intelligibility. *The Modern Language Journal, 106*(4), 744-763.

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: a meta-analysis. *Applied Linguistics, 36*(3), 345-366.

Levis, J. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly,* 39(3), 369-377.

Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, 6(3), 310-328.

Lin, L. (2009). *A study of using 'shadowing' as a task in junior high EFL program in Taiwan*. [Unpublished master's thesis]. National Taiwan University of Science and Technology.

Macaro, E. (2019). Systematic reviews in Applied Linguistics. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in Applied Linguistics* (pp. 230-239). Routledge.

Macaro, E., Handley, Z., & Walter, C. (2012). A systematic review of CALL in English as a second language: focus on primary and secondary education. *Language Teaching, 45*(1), 1-43.

Macaro, E., Curle, S., Pun, J., An, J., & Dearden, J. (2018). A systematic review of English medium instruction in higher education. *Lang. Teach., 51*(1), 36-76.

Macdonald, S. (2002). Pronunciation: views and practices of reluctant teachers. *Prospect,* 17(3), 3-18.

Major, R. (2008). Transfer in second language phonology: a review. In J. Hansen Edwards & M. Zampini (Eds.), *Phonology and second language acquisition* (pp. 63-94). John Benjamins.

Martinsen, R., Montgomery, C., & Willardson, V. (2017). The effectiveness of video-based shadowing and tracking pronunciation exercises for foreign language learners. *Foreign Language Annals, 50*(4), 661-680.

Mishima, M., & Cheng, L. (2017). The impact of a computer-mediated shadowing activity on ESL speaking skill development: a pilot study. *L2 Journal, 9*(1), 21-35.

Mochizuki, H. (2006). Application of shadowing to TEFL in Japan: the case of junior high school students. *Studies in English Language Teaching*, *29*, 29-44.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, 4(1). https://doi.org/10.1186/2046-4053-4-1

Mori, Y. (2011). Shadowing with oral reading: effects of combined training on the improvement of Japanese EFL learners' prosody. *Language Education & Technology, 48*, 1-22.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. *TESOL Quarterly, 25*(3), 481-520.

Murphy, D. (2011). An investigation of English pronunciation teaching in Ireland. *English Today, 27*(4), 10-18.

Murphy, J. (2014). Myth 7: teacher training programs provide adequate preparation in how to teach pronunciation. In L. Grant (Ed.), *Pronunciation myths: applying second language research to classroom teaching* (pp. 188-224). University of Michigan Press.

Nakanishi, Y., & Nakanishi, Y. (2015). Use of an intermediate face between a learner and a teacher in second language learning with shadowing. *AH'15: Proceedings of the 6th Augmented Human International Conference,* 113-116.

Nakayama, T. (2021). Effectiveness of the visual-auditory shadowing method on learning the pronunciation of kanji. *Japanese Psychological Research, 63*(1), 26-34.

Nakayama, T., & Armstrong, T. (2011, November). *Weak forms in shadowing: how can Japanese EFL learners perform better on shadowing tasks?* [Paper presentation]. Japan Association for Language Teaching (JALT) International Conference, Tokyo.

Newman, M., & Gough, D. (2020). Systematic reviews in educational research: methodology, perspectives and application. In P. Zawacki-Richter, M. Kerres, S. Bedenlier, M. Bond, & K. Buntins (Eds.), *Systematic reviews in educational research: methodology, perspective and application* (pp. 3-22). Springer.

Nguyen, L., & Newton, J. (2020). Pronunciation teaching in tertiary EFL classes: Vietnamese teachers' beliefs and practices. *TESL-EJ, 24*(1), 1-20.

Nguyen, H., & Nguyen, M. (2019). Applying shadowing technique and authentic materials to promote phonological awareness amongst young learners of English. *Proceedings of ELT Upgrades 2019: a focus on methodology*, 13-23.

Omar, H., & Umehara, M. (2010). Using a "shadowing technique" to improve English pronunciation of deficient adult Japanese learners: an action research on expatriate Japanese adult learners. *The Journal of Asia TEFL, 7*(2), 199-230.

Ono, Y., Ishihara, M., & Yamashiro, M. (2012). Mobile-based shadowing materials in foreign language teaching. *The 1st Global Conference on Consumer Electronics 2021*, 90-93.

OxfordHouse. (2018, April 9). *Shadowing: a new way to improve fluency at C2 level*. https://oxfordhousebcn.com/en/shadowing-improve-fluency-at-c2-level/

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: a practical guide.* Blackwell Publishing.

Pourhosein Gilakjani, A. (2012). A study of factors affecting EFL learners' English pronunciation learning and strategies for instruction. *International Journal of Humanities and Social Science, 2*(3), 119-128.

Pronunciation with Emma. (2022, September 2). *British English pronunciation: shadowing exercises (listen and repeat)* [Video]. YouTube. https://www.youtube.com/watch?v=-Lf2E_tsuCM

Rau, D., Chang, A., & Tarone, E. (2009). Think or sink: Chinese learners' acquisition of the voiceless interdental fricative. *Language Learning, 59*, 581-621.

Rose, H., Briggs, J., Boggs, J., Sergio, L., & Ivanova-Slavianskaia, N. (2018). A systematic review of language learner strategy research in the face of self-regulation. *System, 72*, 151-163.

Rojczyk, A. (2013). Phonetic imitation of L2 vowels in a rapid shadowing task. In J. Levis & K. LeVelle (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference.* Aug. 2012. (pp. 66-76). Iowa State University.

Rongna, A., & Hayashi, R. (2012). Accuracy of Japanese pitch accent rises during and after shadowing training. *Proc. Speech Prosody 2012*, 214-217.

Rongna, A., Hayashi, R., & Kitamura, T. (2013). Naturalness on Japanese pronunciation before and after shadowing training and prosody modified stimuli. *SLaTE 2013*, 143-146.

Rongna, A., Hayashi, R., & Kitamura, T. (2015). Crucial prosodic features in Japanese learners' pronunciation: evidence from naturalness judgments of synthetic speech. *Journal of the Phonetic Society of Japan, 19*(3), 37-42.

Rose, H., McKinley, J., & Galloway, N. (2021). Global Englishes and language teaching: a review of pedagogical research. *Language Teaching, 54*(2), 157-189.

Saito, K. (2012). Effects of instruction on L2 pronunciation development: a synthesis of 15 quasi-experimental intervention studies. *TESOL Quarterly, 46*(4), 842-854.

Saito, K., & Brajot, F. (2013). Scrutinising the role of length of residence and age of acquisition in the interlanguage pronunciation development of English /r/ by late Japanese bilinguals. *Bilingualism: Language and Cognition, 16,* 857-863.

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.

Saito, Y., Nagasawa, Y., & Ishikawa, S. (2011). Effective instruction of shadowing using a movie. In A. Stewart (Ed.), *JALT2010 Conference Proceedings* (pp. 139-148). JALT.

Shah, S., Othman, J., & Senom, F. (2017). The pronunciation component in ESL Lessons: teachers' beliefs and practices. *Indonesian Journal of Applied Linguistics*, *6*(2), 193-203.

Shao, Y., Saito, K., & Tierney, A. (2023). How does having a good ear promote instructed second language pronunciation development? Roles of domain-general auditory processing in choral repetition training. *TESOL Quarterly, 57*(1), 33-63.

Shemesh, H. (2022, February 22). *Shadowing technique in English: are you wasting your time?* Accent's Way Magazine. https://hadarshemesh.com/magazine/shadowing-in-english/

Shiki, O., Mori, Y., Kadota, S., & Yoshida, S. (2010). Exploring differences between shadowing and repeating practices: an analysis of reproduction rate and types of reproduced words. *Annual Review of English Language Education in Japan, 21*, 81-90.

Šturm, P., Przedlacka, J., & Rojczyk, A. (2022). Phonetic imitation of t-glotalling by Czech speakers of English. *Linguistica Pragensia, 1*, 142-165.

Sumiyoshi, H. (2014). *Exploring the effects of the shadowing method: case studies of Japanese language learners at an Australian university* [Unpublished doctoral dissertation]. Macquarie University.

Sumiyoshi, H., & Svetanant, C. (2017). Motivation and attitude towards shadowing: learners' perspectives in Japanese as a foreign language. *Asian-Pacific Journal of Second and Foreign Language Education, 2*(16). https://doi.org/10.1186/s40862-017-0039-6

Talebzadeh, H., & Gholami, L. (2015). The relationship between English pronunciation self-concept and English learning. *International Letters of Social and Humanistic Sciences*, 60, 54-59.

Tamai, K. (1992). Follow-up no chokairyoku kojo ni oyobosu koka oyobi "follow-up" noryoku to chokairyoku no kankei. dai 4 kai "Eiken" kenkyu josei hokoku [The effect of follow-up on listening comprehension]. *STEP Bulletin*, 4, 48–62.

Tamai, K. (1997). Shadowing no koka to chokai process ni okeru ichizuke [The effectiveness of shadowing and its position in the listening process]. *Current English Studies,* 36, 105–16.

Teeter, J. (2017). Improving motivation to learn English in Japan with a self-study shadowing application. *Languages, 2*(19). https://doi.org/10.3390/languages2040019

Thomson, R. (2017). Measurement of accentedness, intelligibility, and comprehensibility. In O. Kang & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 11-29). Routledge.

Thomson, R., & Derwing, T. (2015). The effectiveness of L2 pronunciation instruction: a narrative review. *Applied Linguistics, 36*(3), 326-344.

Varasarin, P. (2007). *An action research study of pronunciation training, language learning strategies and speaking confidence*. [Unpublished doctoral thesis]. Victoria University. https://vuir.vu.edu.au/1437/3/VARASARIN%20Patchara-thesis_nosignature.pdf

Vitanova, G., & Miller, A. (2002). Reflective practice in pronunciation learning. *The Internet TESL Journal, VIII* (1). http://iteslj.org/Articles/Vitanova-Pronunciation.html

Wahid, R., & Suhlong, S. (2013). The gap between research and practice in the teaching of pronunciation: insights from teachers' beliefs and practices. *World Applied Sciences Journal Special Issues of Studies in Language Teaching and Learning*, 21, 133-142.

Wang, X. (2018). The study of shadowing exercise on improving oral English ability for non-English major college students. *Advances in Social Science, Education and Humanities Research, 120*, 195-200.

Willardson, V. (2014). *The effectiveness of computer-enhanced shadowing and tracking pronunciation exercises for intermediate level foreign language learners* [Unpublished master's dissertation]. Brigham Young University.

Yates, L. (2011). Language, interaction and social inclusion in early settlement. *International Journal of Bilingual Education and Bilingualism, 14*(4), 457-471.

Yavari, F., & Shafiee, S. (2018). Effects of shadowing and tracking on intermediate EFL learners' oral fluency. *International Journal of Instruction, 12*(1), 869-884.

Zając, M., & Rojczyk, A. (2014). Imitation of English vowel duration upon exposure to native and non-native speech. *Poznán Studies in Contemporary Linguistics, 50*(4), 495-514.

Zajdler, E. (2020). Speech shadowing as a teaching technique in the CFL classroom. *Lingua Posnaniensis, 62*(1), 77-88.

Zawacki-Richter, O., Kerres, M., Bedenlier, S., Bond, M., & Buntins, K. (2020). *Systematic reviews in educational research: methodology, perspective and application.* Springer.

Zhang, K., & Peng, G. (2017). The relationship between the perception and production of non-native tones. *Interspeech 2017*, 1799-1803.

Zhang, X., Miyaki, T., & Rekimoto, J. (2016). WithYou: an interactive shadowing coach with speech recognition. *UIST'16 Adjunct*, 61-63.

Zhang, X., Miyaki, T., & Rekimoto, J. (2020). WithYou: automated adaptive speech tutoring with context-dependent speech recognition. *CHI'20*. https://doi.org/10.1145/3313831.3376322

Zielinski, B., (2011). The social impact of pronunciation difficulties: confidence and willingness to speak. *Pronunciation in Second Language Learning and Teaching Proceedings,* 3(1).

Zoss, J. (2015). *Adult English learners' perceptions of their pronunciation and linguistic self-confidence*. [Unpublished master's thesis]. Hamline University. https://core.ac.uk/download/pdf/230676143.pdf

Zoss, J. (2016). What do adult English learners say about their pronunciation and linguistic self-confidence? *MinneTESOL Journal, 32*(2). http://minnetesoljournal.org/wp-content/uploads/2018/10/ZossFall2016MinneTESOLJournal.pdf

Annex 1.

REVIEW PROTOCOL

Section 1. Administrative information

Title
Shadowing for pronunciation: a systematic review.

Registration
This protocol has not been registered.

Author
Benen Nadelek Whitworth, University of Oxford
MSc student in Applied Linguistic for Language Teaching

Contributions
Benen Nadelek Whitworth (BW) is the author of this review, and will undertake the development of the review, study selection and data analysis.
Filip Bigos (FB), an MSc student on the same course, is a second reviewer and will contribute to title and abstract screening for study selection.

Support
No funding has been received to finance this review.

Section 2. Introduction

Rationale
Intelligible, comprehensible pronunciation is essential for clear and effective communication in an L2. However, pronunciation teaching is often under-emphasised in teacher training and classroom practice, and teachers often feel doubtful of how to teach pronunciation effectively.

For these reasons, it is of great importance to research innovative pronunciation teaching techniques. Shadowing is one such technique currently becoming very popular worldwide. This review will explore whether shadowing is beneficial in improving student pronunciation, and what students think of the technique.

Objectives
To explore research on shadowing for pronunciation, this review aims to respond to the following research questions:

Over-arching RQ) What is the extent and nature of research on the use of shadowing in teaching pronunciation in L2 classrooms?

RQ1) What evidence is available on the effectiveness of shadowing to improve learners' pronunciation?

RQ2) How do learners evaluate this teaching and learning technique?

Section 3. Methods

Eligibility criteria

The review will employ the following eligibility criteria:

1) Empirical evidence: the studies included must provide empirical evidence.
2) Year of publication: the studies included must be published after 1992.
3) L2 learning: the studies included must investigate shadowing for L2 language learning.
4) L2 pronunciation: the studies included must investigate shadowing for L2 pronunciation development.

Information sources

Databases used will be the following: Web of Science, Scopus, the British Education Index, the Australian Education Index, and the Educational Resources Information Centre (ERIC). These databases will provide access to studies published in peer-reviewed journals.

ProQuest Social Science Premium Collection will also be used to provide access to unpublishes dissertations and theses.

The search will be conducted in June 2023.

Search strategy

The search strategy used will include the following terms:

	Shadowing	**Pronunciation**	**Second language learning**	**Filters**
Place in article	Abstract	Abstract	Abstract	All document
Term(s)	Shadowing	AND Pronunciation OR Speaking OR Oral OR Speech	AND Learn* OR Stud* OR Language*	NOT Medicine OR Dentistry OR Nursing OR Ultrasound OR Cocktail

				OR Satellite

Exact search strings will be adapted to each data base, and documented accordingly.

Data management
The process of study selection will be managed by the website Rayyan. https://rayyan.ai/
References for studies will be handled manually.

Selection process
BW will use the eligibility criteria to undertake title and abstract screening of all identified studies. FB will undertake second screening for titles and abstracts of 10% of studies, to calculate interrater reliability. Screening at this stage will be overinclusive.

BW will then conduct full text screening to select the final sample of studies.

Data collection process
A data extraction grid, developed by BW in coordination with her MSc supervisor, will be used to collect data for this review.

Data items
The following data items will be extracted from every study:

- Author
- Title
- Year of publication
- Study type (e.g. journal article, dissertation, etc.)
- Database
- Quality of abstract
- Quality of introduction
- Quality of literature review
- Research questions
- Methodology used (e.g. quantitative, qualitative, mixed)
- Research design (e.g. case study, experiment, action research, etc. and cross-sectional or longitudinal)
- Variables
- Sampling technique
- Sample size
- Sample characteristics (age, gender, L1, L2, proficiency, educational level, country)
- Data collection instruments
- Validity and reliability or trustworthiness of data collection instruments
- Whether tasks used are in line with definitions of shadowing
- Type of data analysis used
- Reliability and validity of data analysis
- Results of study
- Conclusions of study
- Whether implications match study findings
- Limitations of study
- Suggestions for further research

Quality assessment and risk of bias in individual studies
Weight of evidence judgments will also be given per study, for 4 criteria: relevance, appropriateness, contribution, trustworthiness. Scores given will be low, medium, or high.

A critical review will also be written for each study, highlighting the studies' key strengths and weaknesses.

Data synthesis
Data will be qualitatively synthesised and organised according to the research questions. That is, according to the effectiveness of shadowing for pronunciation, and how students evaluate the technique.

Annex 2.

EXTRACTION GRID FOR SYSTEMATIC REVIEW ON SHADOWING FOR IMPROVING PRONUNCIATION

Reviewer:
Study ID:

General information about study	
Author	
Title	
Year	
Paper type	
Database	

In-depth data extraction		
In-depth review item	**Criteria for quality**	**Reviewer response: yes/no/partly/unclear/brief comment**
Abstract	Does the abstract provide sufficient information to understand the study?	
Introduction/rationale	Does the introduction report the topic of the study clearly?	
	Does the introduction explain why the study was carried out?	
Literature review	Does the literature review define shadowing?	
	Does the literature review link to theoretical frameworks to understand shadowing?	
	Does the study report any other core constructs?	

Research questions	Are RQs clearly stated?	
Methodology	What, broadly, is the methodology used? (quant/qual/mixed)	
	What is the research design? (e.g. experiment, case study, survey, action research, ethnography, etc.)	
	Was the study cross-sectional or longitudinal? Please specify (e.g. length of study)	
Variables	(If applicable) Is it clear what the dependent and independent variable(s) was/were? Please specify if so.	
	What other variables, if any, are controlled for?	
	Is the population clearly described? Please specify.	
Sampling	Does the study report the planned sample size and characteristics?	
	Does the study report the actual sample size and characteristics? Please specify.	
	Is a sampling strategy provided? Please specify.	
	Does the study report the participants' age? Please specify.	
	Does the study report the participants' gender? Please specify.	
	Does the study report the participants' L1? Please specify.	

	Does the study report the participants' L2? Please specify.	
	Does the study report the participants' L2 proficiency? Please specify.	
	Does the study report the participants' educational level? Please specify.	
	Does the study report the country/countries where the study was carried out? Please specify.	
	Are there any other important features of the participants/research context?	
Data collection	What data collection instruments were used?	
	(If quantitative) Does the study report how they addressed the validity and reliability of their data collection instruments?	
	(If qualitative) Do data collection instruments seem trustworthy?	
	Did the data collection seem reliable?	
	Are tasks used for data collection in line with established definitions of shadowing? Please specify if not.	
Data analysis	Does the study report how data were analysed? Please specify.	
	Does this seem to be a reliable/valid method of analysing the data?	

	Does the analysis match the requirements of the research questions?	
Results, findings, discussion, and conclusion	What are the actual results of the study?	
	What does the study conclude about these results?	
	Do conclusions match your assessment of the results?	
	Does the study discuss implications for using shadowing to improve pronunciation? Please specify.	
	Do the implications match the study findings?	
	Are limitations of the study discussed?	
	Are there suggestions for further research? Please specify.	

Initial weighting by reviewer	
Importance of article to the review	**High/medium/low**
WOE: RELEVANCE. Does the focus of this study match the objective of this specific review? (Is the topic relevant?)	
WOE: APPROPRIATENESS. Is the design and analysis suitable in addressing the questions, or sub-questions of this specific review? (Is the design relevant?	

WOE: CONTRIBUTION. Does the study answer the question of this specific review and help inform use of shadowing to improve pronunciation? (Does it add to knowledge?)	
WOE: TRUSTWORTHINESS. Taking account of all quality assessment issues, can the study findings be trusted to answer its research questions? (Can we trust its findings?)	

Critical review of article

Annex 3.

CRITICAL REVIEWS OF ALL INCLUDED STUDIES

1)

Althubyani, R. (2021). ***The effect of shadowing in learning L2 segments: a perspective from phonetic convergence.*** **[Unpublished master's thesis]. University of Wisconsin Milwaukee.**

Althubyani (2021) conducted a study to explore the effect of shadowing training on production of the English segmentals /b/, /v/, /ɛ/ and /oʊ/ amongst Arabic speakers. The study also investigated whether participants showed more improvement in pronunciation when trained by a native model talker they phonetically converge to, and whether pre-existing phonetic distance between model talkers and shadowers played a role in the degree of convergence.

Participants were 26 female Saudi native Arabic speakers studying in the USA. They were aged between 20 to 47 (M = 28.42, SD = 6.12) and their length of residency in the USA ranged from 2 to 10 years (M = 5.27, SD = 2.46). All had studied English as foreign language from around the age of 13 in Saudi Arabia, and all had studied English in the USA for at least two semesters. Their IELTS scores ranged from 5 to 6.5, indicating an intermediate level of proficiency. Participants were recruited from 3 different US universities.

First, a production pre-test was conducted to measure participants' baseline production of a list of 12 English words. The words were presented in 5 different blocks with additional filler words to avoid over-familiarising learners with the target words. Target words selected were disyllabic, as phonetic convergence is more evident in such words, and low frequency, to evoke more convergence. All words were stressed on the first syllable with a syllable structure of CVCVC or CVCCVC to align with how stress works in words in Arabic language and thus helping participants produce the vowels in the same manner as native speakers. The first syllables in all words had vowels /i/, /æ/, /ʌ/, which were selected based on the fact that Saudi speakers of English often show most convergence in these sounds.

Next, participants shadowed five model talkers, all female speakers of Upper Midwestern American English, saying the same list of words and also filler words. Learners shadowed the five model talkers in five randomised blocks on the same day.

Phonetic convergence between the participants and the model speakers was measured through perceptual similarity judgement tasks and acoustic data. 5 perceptual similarity judgement tasks were used, and involved 55 English-speaking monolinguals born in the USA and recruited from the platform Prolific (11 listeners per task). No listeners had speaking or hearing impairments. The tests involved selecting which production from a given participant (baseline or shadowing) sounded most similar to the model talker. The presentation of baseline and shadowing productions was counterbalanced. The task took 45-60 minutes, and four attention checks were added to the task to make sure that listeners were paying attention. Data from listeners who missed the attention checks was excluded from the study. Inter-rater reliability was high for some of the tasks (tasks 1 and task 5), lower for others (task 3 and task 4), and poor for one (task 2). Statistical analysis, involving one-sample t-tests for parametric data and a one-sample Wilcoxon signed rank test for non-parametric data, suggested that listeners chose

shadowed items as closer to the model items than the baseline in all 5 tasks. A Kruskal-Wallis test indicated that there were differences in perceived convergence across the five model talkers.

Acoustic measurements, via Praat software, were also used to measure convergence. This included values of VOT, vowel duration, F0, and F1*F2. These values were converted to difference-in-distance (DID) scores, which estimated the effect of model talker on shadowed productions by comparing baseline differences between each participant and each model talker, as well as shadowed differences. Results from analysis of formant values and vowel duration suggested that learners converged the majority of their vowels to model talkers, but the degree of convergence varied. However, convergence for VOT of stops and F1*F2 did not reach a significance. Other analyses suggested that the greater the pre-existing baseline distance between participants' and model speakers' productions, the more the learners converged.

Based on the degree of convergence among learners and native model speakers, as measured by perceptual judgment and acoustic features, 20 learners were assigned to two equal training groups. One group (convergence or C-group) was trained by the model talkers to whom they showed the highest degree of phonetic convergence, and the other (divergence or D group) received training from model talkers who they diverged from or showed the least convergence to. Each learner was trained by one native model talker, in a low-variability paradigm, to fully explore the phenomenon of convergence.

After matching trainees and trainers, trainees learnt 12 monosyllabic English nonsense words divided into two sets, one of which included the target consonants /p/ and /v/, and the other of which included the target vowels /ɛ/ and /oʊ/. The words were designed so that no features other than the target sounds would present difficulties for Arabic speakers, e.g., syllable structure.

In the pre-test, audio recordings were made of the trainees producing the nonsense words after hearing them and seeing a non-object presented with them on a screen. They were asked to repeat as quickly and clearly as possible. Participants then undertook a three-day training procedure, with the same wordlist was used in all three sessions. Each word was repeated 12 times in each session. Activities in all training sessions was the same, with the only feature differing in the sessions being the audio, i.e., the voice of the model speaker. Finally, participants undertook a post-test directly after their last training session. The post-test was identical to the pre-test, to explore how much pronunciation of the target segments had improved.

After post-test, 40 native English listeners completed 4 judgement tasks (10 listeners per task) to assess the intelligibility of each recording. No listeners had hearing or speaking impediments, and all were monolingual English speakers born in the USA. Listeners who failed attention checks were excluded from the study. Each judgement task assessed a different target feature: /p/, /v/, /ɛ/ and /oʊ/. In all tasks, listeners identified participants' productions as either the target sound, a different but similar sound (e.g., /b/ for /p/, /f/ for /v/) or neither. Inter-rater reliability was calculated for all tasks and was high.

Various statistical tests were carried out to explore the results. A significant main effect was found for time, meaning that both groups showed an increase in intelligibility scores across the two time periods. However, there was no main effect for groups, suggesting no difference between the two group's intelligibility at pre- and post-test. In addition, although C-group

showed more intelligible segments after training than D-group, and a greater magnitude of change, this difference was not statistically significant.

For production of /p/, both groups improved significantly from pre- to post-test, but there was no significant main effect for group. For production of /v/, there were no significant improvements for either group from pre- to post-test, and no significant differences were found between groups at pre- or post-test. For production of /ɛ/, there was a significant main effect for time, with both groups doing significantly better in the post-test. There was no main effect for groups. For production of /oʊ/, there was no significant difference between the groups at pre-test, but there was at post-test, with C-group doing significantly better than D-group after training. There was also a significant increase in intelligibility from the pre-test to post-test for both groups. Overall, results suggested that both groups showed significant improvement from pre- to post-test in all target segments except /v/, due to the high performance on this sound at pre-test.

GEE models confirmed that perceived phonetic convergence was not a significant predictor of improvement. However, it did suggest that, regardless of group, convergence in vowel duration and vowel spectra, as measured by acoustic data, were significant predictors of trainees' performance. In addition, the larger the pre-existing distance in a given phonetic dimension was, the greater the degree of convergence.

Overall, the author concludes that shadowing is effective in improving segmental intelligibility. In addition, whilst allocation to convergence or divergence groups did not explain variation in performance, shadowing speakers whose voices were similar to those of L2 learners in terms of vowel duration and vowel spectra resulted in more improvement.

The study is rigorous in its explanation of data collection instruments, which appeared valid, and reliability of analysis. It also provides a rigorous explanation of the procedure through which participants were assigned to the convergence or divergence group. Statistical analyses seem sound and results appear trustworthy. The authors also acknowledge two key limitations of the study. Firstly, its questionable ecological validity due to the fact that lab settings often minimise the difficulties encountered by L2 speakers, and the fact that the words selected to measure convergence were selected so as not to pose a difficulty to participants. Secondly, the heterogeneous but small sample, making it very difficult to control for factors like length of residency and age.

It should be noted, however, that despite the trustworthiness of its findings, this study does not employ tasks in line with shadowing, as defined by this review. "Shadowing" in the study is, effectively, repetition of lists of words as simultaneously as possible. As such, this task does not meet the criteria of speaking and listening simultaneously to incoming parts of the same utterance, and was not included for discussion in the results section.

2)

Bovee, N., & Stewart, J. (2008). The utility of shadowing. In A. Stoke (Ed.), *JALT2008 Conference Proceedings* (pp. 888-900). JALT.

Bovee and Stewart (2008) investigated the effect of shadowing on student pronunciation. To do so, they introduced shadowing as a weekly homework activity for 400 Japanese university students. Age, gender, and L1 of participants is not reported, but participants are assumed to

largely be young adult Japanese speakers. Each week, students had to complete a recorded shadowing activity and send these recordings to their teachers.

To analyse improvement from week 1 to week 10, 21 students' recordings from these two weeks were randomly selected. 8 native English speaking raters were asked to judge which recording (week 1 or week 10) represented more natural English pronunciation. The order in which raters heard the recordings was random, and they were not able to select that both recordings were of the same quality. Raters were unable to rate both files as of the same quality.

Two different levels of interrater agreement were explored in order to define "improvement". According to the less stringent improvement criterion (5/8 raters agreed), 83% of lower-level students (TOEIC Bridge score 0-100), 50% of intermediate students (TOEIC Bridge score 101-120), and 86% of higher-level students (TOEIC Bridge score above 131) improved from the recording in week 1 to week 10. According to the more stringent improvement criterion (6/8 raters agreed), 67% of lower-level students, 50% of intermediate students, and 29% of higher-level students improved from recordings in week 1 to week 10.

As well as the assessment of the recordings, the authors also conducted a survey of 89 students' perceptions of their own improvement and of shadowing in general. 86% of students believed shadowing improved their listening skills in the intervention, 67% believed they improved their pronunciation of individual words, 73% their overall intonation, and 80% believed shadowing had overall value. There are also breakdowns given per student proficiency level for the survey, in which lower-level students tended not to believe they improved their pronunciation of individual words (38%) or overall intonation (56%), in contrast to what raters indicated.

As the authors indicate, the study could be improved by less subjective, more fine-grained measures of what constitutes "good pronunciation", rather than the intuitive, holistic assessment used. Such a scale should ideally include measures less related to "native-like" pronunciation, and more related to comprehensibility, in line with current pronunciation theory and perspectives from the field of Global Englishes. In addition, a more precise measure of pronunciation would also allow raters to justify their decisions when both recordings appeared to be of similar quality. This would also improve trustworthiness of results, as improvements could be somewhat inflated in the current study due to raters being unable to specify when recordings were of the same quality. In addition, rather than only using shadowing recordings to measure pronunciation improvement, a spontaneous task could be used to capture and analyse more ecologically valid speech samples.

In terms of results related to student perceptions, it is unclear how and by whom the surveys were administered. If surveys were administered by teachers, it is possible that perceptions reported may be overly positive, as students aimed to please their instructors or feared negative consequences for critical responses.

3)

El-Esery, A. (2021). English suprasegmentals for EFL learners: an experimental study with shadowing technique. ***Asian EFL Journal Research Articles, 28*****(3.2), 54-71.**

El-Esery conducted a 3-month shadowing intervention with 35 English learners on an intensive course at Qassim University, Saudi Arabia. The objective of the study was to explore whether the intervention improved participants' listening ability and production of intelligible speech.

To measure improvement, two pre- and post-tests were conducted. The first was a listening comprehension test, and the second was a listening-shadowing test, in which participants were asked to listen to audio extracts with transcripts showing various gaps. To fill the gaps, participants had to shadow reduced forms of words and complete their written forms. In the listening-shadowing test, participants received one point for writing the full form of words and phrases, and one point for shadowing correctly. Both tests were submitted to a panel of language instructors and EFL specialists at the university for feedback on the validity and feasibility of the tests. Cronbach's alpha was also calculated, and suggested high internal consistency for both tests (0.87 for the listening comprehension test, 0.813 for the shadowing test). The versions of the test appear to have been identical at pre- and post-test. Whilst the scoring method is described in detail, the number of raters and calculations of interrater reliability are not provided.

To analyse the results of the pre- and post-tests, paired samples t-tests were used. Results showed that mean scores on the listening test improved from 30.11 to 39.0, which was found to be statistically significant at 0.001. Mean scores on the shadowing-listening test improved from 16.6 to 21.6, which was also found to be significant at 0.001.

The author concludes that these results demonstrate that shadowing can be effective in improving listening comprehension and speech intelligibility. Whilst the first claim seems reasonable, the second is problematic for several reasons. Firstly, "speech intelligibility" is not defined in the paper, and is measured by a composite score of identifying suprasegmental features and successful shadowing. This does not appear to be a valid way of measuring the construct. There are anecdotal claims that participants improve in their oral use of suprasegmentals in a reflection stage of the shadowing cycle, but this is not developed in any meaningful way. However, the paper does provide evidence to suggest that participants' awareness and production of weak forms improved during training.

In addition, the generalisability of the results is limited by opaque reporting: no explanation is given about the exact shadowing materials or procedure used, or about certain aspects of the system of rating participants' speech in the listening-shadowing test (raters, scoring system, or inter-rater reliability). The lack of a control group also limits the findings. As such, whilst the paper presents useful data on the impact of shadowing, the data it presents on improvements in pronunciation should be taken with caution.

4)

Foote, J., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. ***Journal of Second Language Pronunciation, 3*****(1), 34-56.**

Foote and McDonough (2017) investigated whether shadowing practice could improve advanced L2 English speakers' pronunciation, and how these participants felt about the intervention.

To do so, they recruited 22 L2 speakers of English who were studying at an English medium university in Montreal, Canada. Of these 22 participants, 16 stayed for the duration of the study. Participants were between 18 to 38 years old, with 7 males and 9 females. They came from different L2 backgrounds (Chinese = 10, French = 3, Arabic = 1, Bengali = 1, Russian = 1), and had been in Canada for between 1 and 60 months.

The shadowing intervention was an 8-week program, in which participants used iPods to practice shadowing one-minute dialogues from popular sitcoms (e.g. Friends, the Big Bang Theory) at least four times a week, for a minimum of 10 minutes each time. To ensure participants were practising, they were required to submit a sample recording via email after each practice session.

Pre-, mid-, and post-tests were administered in weeks 1, 6, and 8, to measure pronunciation improvement. The tests included a picture narrative task, and a shadowing task similar to the dialogues participants were already practising with, both of which were recorded. The picture narrative task appears to have been the same at all time points, but it is unclear whether the shadowing task varied.

Recordings were assessed by 22 English L1 raters, who were recruited from a university in Alberta, Canada. They were not professional raters, but were trained to use a computer-based rating system on MATLAB. Raters listened to all recordings from the picture narrative task and rated each speech sample for accentedness, comprehensibility, and fluency on three different scales (ranging from 1-1000). They also listened to the shadowing task recordings and rated how well the L2 speakers were able to imitate the speech model. Raters heard recordings in a randomised order, and were not aware of which time point the recordings were from. Inter-rater reliability was calculated using intraclass correlation coefficients, and a high degree of agreement was found on all measures: shadowing ($\alpha = .86$), accentedness ($\alpha = .91$), comprehensibility ($\alpha = .89$), and fluency ($\alpha = .93$).

Results showed that, overall, participants showed significant improvement on comprehensibility ($p = .01$), shadowing ($p = .0001$), and fluency ($p = .0001$), although not on accentedness ($p = .05$). It should be noted that, although there was overall improvement on these measures, there was not always improvement from one testing time to the next, indicating potential minimum thresholds for improvement or possible points at which the benefits of shadowing decline.

In addition to the pronunciation pre-, mid- and post-tests, structured interviews with a range of 9-point Likert-scale and open-ended questions were conducted in weeks 1, 6, and 8. Participants seemed to enjoy shadowing more and value it more as an intervention to improve pronunciation from time 2 to time 3 (scores increased from 7.5 to 7.63 on a 9-point Likert scale). They also seemed to be more convinced it was helping their pronunciation by the end of the intervention (6.81 at time 2, 7.5 at time 3). It should be noted that statistical significance is not calculated for these changes. In open questions about overall views on the shadowing project, all but one responses were positive. All participants answered that they thought shadowing was an effective way to improve pronunciation. 10 participants believed their pronunciation had improved, 1 thought it hadn't, and 5 gave qualified responses. All participants responded that they would recommend shadowing to friends and family.

This study is notable for its inclusion of both controlled and spontaneous tasks to gather speech samples, and its careful description of materials, procedures, raters, and rating procedures used. Its separation of the constructs of accentedness, fluency, and comprehensibility into different scales also allows for more meaningful results. As such, it is one of the most robust and appropriate studies in this review.

However, there are several issues which temper its findings. As the authors note, the fact that 6 participants left the study may have introduced an element of participant self-selection into the project, and thus have skewed the interview data. They also note the issue of the lack of control group, although they make the case that use of the picture-based task used in other research with immigrants' pronunciation adds comparability. However, this issue is crucial as some participants were new arrivals in Canada and L2 pronunciation improvement in adults often takes place in the first year of exposure in naturalistic environments (Flege, 1988; Munro & Derwing, 2008). Another notably methodological issue is the use of identical tests over three time points, which may have contributed to improvements, particularly in fluency.

A final feature of the study not addressed by the authors is the fact that the majority of participants were L1 Chinese speakers (n = 10), 8 of whom had been in Canada for less than 6 months. This could have influenced results in various ways. Firstly, speakers of languages with a similar sound and prosody system to English, such as German, could, for instance, have shown less improvement. Secondly, shadowing may have been particularly effective for students who had not yet had a great deal of contact with everyday spoken English, and had come from a radically different L1 background.

5)

Guo, S., & Hsu, T. (2017). Integration of peer assessment and shadowing strategies for improving the oral performance of EFL learners. In W. Chen et al. (Eds.), ***Proceedings of the 25th International Conference on Computers in Education*** **(pp. 928-230). Asia-Pacific Society for Computers in Education.**

Guo and Hsu (2017) investigated the impact of shadowing training on students' oral ability in English.

Participants were 57 Taiwanese English majors, aged 17. No information is given on how participants were recruited, nor on their L1, English proficiency, or gender. Participants were divided into a control (n = 30) and experimental (n = 27) group. An unspecified pre-test measure of oral ability showed no significant differences between the groups. It should be noted that it is not stated whether the two groups were taught by the same teacher or not. Both groups submitted recordings of themselves responding to 2 sets of oral questions, taken from the General English Proficiency Test and reviewed by three English teachers to ensure they were suitable for the study. The control group responded directly to these questions, whilst the experimental group shadowed audio input selected by teachers from the internet prior to answering the questions.

The recordings from the questions were scored for oral ability. It should be noted that no information is given as to the scoring rubric, and, importantly for this review, whether it included measures of pronunciation or not. From information found online, the General English Proficiency test is a reliable and validated test used in Taiwan, that uses a scoring system that encompasses elements of pronunciation, intonation, and fluency, but also content relevance, and usage of grammar and vocabulary (Language Training and Testing Centre, 2016). The composite nature of the scale, and the fact that the authors do not report specific pronunciation-related scores, makes it difficult to disentangle the various components from one another.

In addition to the lack of information on scoring, no information is given as to who rated the recordings, whether multiple raters were used, and whether raters were blind with respect to whether recordings were from the control or experimental group.

After scoring, t-tests were used to compare scores from the control and experimental groups across time periods. Means scores were also calculated for each group at each time period. Results indicated that there was a statistically significant difference between groups on both tasks, with the experimental group significantly outperforming the control group. The authors conclude from this data that speaking practice with shadowing can help students improve their oral ability and that, through imitating materials with a native-like quality, shadowing allows them to articulate better in their oral production.

Whilst the results do suggest that shadowing helped improve oral ability, the many issues with the study make it difficult to trust the findings. Also, no data is given to support the hypothesis of shadowing having helped with articulation.

Overall, the issues in this study give it a low score on trustworthiness. It also has a low score on relevance as no information is provided specifically on pronunciation.

6)

Hamada, Y. (2018). Shadowing for pronunciation development: haptic-shadowing and IPA-shadowing. *The Journal of Asia TEFL, 15*(1), 167-183.

Hamada (2018) carried out two experiments to explore the impact of two different shadowing techniques on learners' pronunciation. The first technique is Haptic Shadowing (H-SH), in which, as well as shadowing an audio track, learners also simultaneously "punch" the most stressed syllable in the most stressed word. The second International Phonetic Alphabet (IPA) Shadowing (IPA-SH), in which shadowing is accompanied by scripts written in IPA.

Experiment 1

In the first experiment, a total of 58 Japanese second-year university students (18-21 years old) were recruited. 29 were assigned to a H-SH group, and 29 to a IPA-SH group. Both groups were similar in terms of motivation, although the H-SH group contained more low-proficiency learners that the IPA-SH group. It should also be noted that H-SH participants were Engineering majors, whilst IPA-SH participants were Health Science majors. The same materials were used for both groups.

First, a pre-test was administered in which participants read four 50-word sentences loaded with features which are challenging for Japanese learners (i.e., /æ, f, v, θ, ð, w, l, ɹ/). After the pre-test, the H-SH group received an explanation of supra-segmental rules, and the IPA-SH group received an explanation of difficult segmental features for Japanese learners.

The shadowing interventions were then carried out for 15 lessons, over one month, with each group. In both interventions, participants worked with passages of 450-550 words from the textbook *Understanding English Across Cultures*. In the first step of the procedure, both groups worked on the vocabulary and content of the passage. In the second step, the two groups worked with a sub-section of the passage (40-60 words) to shadow. The H-SH group worked on identifying and checking suprasegmental features with their instructor, and listened and

practised reading the sub-section of the passage. They then did silent H-SH, several rounds of vocalised H-SH, the finally recorded their H-SH. The IPA-SH group, however, had a short introduction to one IPA phoneme and how to produce it, then transcribed the sub-section of the passage in IPA. Next, they listened and practised reading, did silent IPA-SH, several rounds of vocalised IPA-SH and script checking, then finally recorded their IPA-SH.

After the intervention, a post-test was administered, in which participants read the same 50-word sentences as in the pre-test.

5 highly proficient non-native speakers were selected as raters to compare pre-test and post-test recordings, and trained so they full understood the rating procedure. The rating procedure was based on intuitive judgments of comprehensibility, segmental features, and suprasegmental features, using a 1-6 Likert scale.

Descriptive statistics showed that both groups improved on comprehensibility, segmentals, and suprasegmentals. However, the H-SH group showed more significant improvement than the IPA-SH group. Inferential statistics nuance this picture slightly, showing that the H-SH group improved significantly on all three features, whilst the IPA-SH group improved only on comprehensibility and segmental features. It should be noted that the fact that H-SH group had more lower proficiency students may have contributed to this difference, as such students could have had more "room for improvement" than their higher-level peers.

Experiment 2

In study 2, Hamada explored student perceptions of the two techniques via the Semantic Differential method factor analysis.

To create enough data for factor analysis, two new groups were added: one receiving H-SH and the other IPA-SH. The extra IPA-SH group included 74 first year Japanese university students studying Health Science, and the extra H-SH group included 75 first year Japanese university students studying Engineering. Each group received their intervention in a procedure very similar to that of study one, with the only difference being that groups in study two practised shadowing for eight lessons, whilst those in study one undertook 15 lessons. After receiving the interventions, all participants from study one and study two were presented with a list of 26 adjectives and asked to select to what degree these adjectives reflected the intervention they had received (on a 1-6 Likert scale). The questionnaire had been tested for reliability in previous work on student perceptions of the standard shadowing method (Hamada, 2011).

Data from the responses of the IPA-SH and H-SH groups was analysed using an exploratory factor analysis. Data was compared between the two groups, and also with student perceptions of standard shadowing as measured by the same questionnaire in previous research (Hamada, 2011). Results showed that learners' impressions of IPA-SH and H-SH were more positive than standard shadowing, as they contained no negative factors. Traditional shadowing, on the other hand, contained five negative factors (useful, shallow, empty, outdated, unpleasant). Overall, students' primary perceptions were of IPA-SH as lively and loud, and of H-SH as useful and enjoyable. H-SH was perceived to be fresher than IPA-SH, being the only type of shadowing to be associated with the item "new". It was also the only type to be associated with the item "favourite".

Whilst the Semantic Differential method does indeed provide a fast way to access student perceptions, it is limited in that it provides pre-assigned adjectives with no way for students to add any other thoughts or descriptions of their own. As such, it may provide an overly simplistic picture of student perceptions, which could best be explored through journals or interviews. It should also be noted that the significant differences between the groups (i.e., differences in proficiency and in major) could have predisposed them to view the interventions differently, limiting the trustworthiness of findings.

7)

Hashimoto, D., Gomi, K., & Shiraishi, R. (2022). Updating L2 phonetic memories: shadowing training vs. listening training. ***Journal of the Phonetic Society of Japan, 26*****, 13-26.**

Hashimoto et al. (2022) explored the effect of shadowing and listening training on convergence towards /p/ and /ɑ/ in 36 target words.

16 Japanese university students (7 males and 9 females) of upper-intermediate proficiency participated in the experiment. Half (5 females, 3 males) were assigned to a listening group and half (4 males, 4 females) were assigned to a shadowing group.

Both groups completed 5 blocks of activities: a warm-up Japanese reading (to familiarise participants with the task and reduce hyper articulation), a pre-training session, a training session, a post-training session, and questionnaires. The only difference in procedure between the two groups was the content of the training session. The listening group completed a training session in which they listened to 36 words carefully, whilst the shadowing group were asked to imitate the same 36 words quickly after hearing them. In pre- and post-training sessions, participants were asked to read the 36 words in a randomised order. Recordings of pre- and post-training sessions were used to analyse Voice Onset Time (VOT) of /p/ and formant values of /ɑ/, to assess the extent to which participants converged to the pronunciation of the target words heard in recordings.

Acoustic measurements were performed using Praat, with annotations carried out by two of the authors and double-checked by the other. Results indicated that, overall, VOT of /p/ was longer in post-training session than pre-training. Formant values for /ɑ/ were also higher in post-training, suggesting that the tongue position was lower and more front in the post-training session. However, when analysed for the effect of group, improvement in VOT of /p/ was only significant in the shadowing group, and improvement in formant values of /ɑ/ was greater in the listening group than the shadowing group.

From these results, the authors conclude that both listening and shadowing training can improve pronunciation. They also suggest that differences in the effectiveness of each method for different sounds may be due to differences in awareness. Listening training may include more time to think, which could be helpful in learning very different vowel sounds which require more awareness. Sub-phonemic differences between /p/ in English and Japanese are very subtle, and may therefore require less awareness and be more trainable through the online nature of shadowing. As each type of training may be suited to training different sounds, the authors suggest both should be used in the classroom, rather than relying solely on one method or the other.

Results from the study are limited by the small sample size involved. In addition, the relevance of the study to this systematic review is highly limited, as what is positioned as "shadowing" appears to be a classical imitation task, with no attention given to simultaneity or listening to incoming speech whilst talking. Indeed, shadowing is not defined at any point in the literature review, and the tasks given and other studies referred to seem to be more in line with imitation tasks than with the shadowing literature. For this reason, the study was not included in the results section of this dissertation.

8)

Haufe, H. (2013). ***A case study of shadowing as a means of helping EAP students to prepare for oral presentations: effects on pronunciation and anxiety.*** **[Unpublished master's dissertation]. Carleton University.**

Haufe (2013) investigated the extent to which 2 weeks of shadowing training helped an EAP student improve pronunciation of target phonemes in oral presentations. She also explored the extent to which shadowing helped with presentation anxiety.

The study involved a case study of one EAP student at Carleton University in Canada. Non-probability, convenience sampling was used to recruit students for the study, with the researcher pitching the project to teachers, one of whom then pitched the projects to their students. Two students volunteered to participate in the study, both of whom were L1 speakers of Mandarin from China. Data from one participant, a 19-year-old male majoring in business, was later excluded from the study, as he did not complete a shadowing log and changed his presentation, so did not have time to practise shadowing for the final version. The data from the other participant, a 21-year-old female majoring in communication, was included, and is the data presented and analysed in the study. The participant had been studying in Canada for two months.

The study took place over 6 weeks and involved eight phases: pre-interview, questionnaire, diagnostic test, written summary, shadowing (2 weeks), presentation rehearsal, in-class presentation, post-interview. The pre-interview and questionnaire were used to determine the participant's previous experience with English, how she viewed her English pronunciation, how familiar she was with shadowing, how much experience she had with giving oral presentations in English, and how comfortable she was presenting in English. The questionnaire was adapted from *The Language Contact Profile* (Freed et al., 2004), to include only items relevant to spoken English and pronunciation. A diagnostic test was used to identify participant's pronunciation problems, and involved use of a reading passage from *Well Said* (Grant, 1993) specifically designed to make students and teachers aware of segmental pronunciation difficulties. Next, the participant was asked to submit a one-page written summary of her presentation topic, which the researcher emphasised should contain core ideas, rather than a word-for-word plan of the presentation, to prevent the participant from memorising what she would later shadow. The participant was given 2 weeks to write this summary. The researcher then edited the participant's summary and recorded herself reading it. After this, the researcher and participant met for a shadowing training session, in which the participant practised shadowing with a passage from *Well Said* (Grant, 1993), selected as its length and academic content were similar to what the participant would shadow to prepare for her oral presentation. The participant was then emailed the recording of her presentation, and instructed to shadow it 20 times over two weeks, and to keep a written log of the date and time she shadowed, and the number of times she shadowed. The day before the in-class presentation,

the participant met with the researcher to give her presentation, which was recorded as a measure of post-shadowing pronunciation improvement, as nerves associated with the in-class presentation were expected to cause more pronunciation mistakes. The in-class presentation was attended by the researcher, and also recorded. Finally, a semi-structured interview was conducted after the presentation to explore the participant's use and assessment of shadowing, and anxiety during the rehearsal and in-class presentation.

Results from the initial questionnaire and interview indicated that, prior to study in Canada, the participant had never studied in any English-speaking country and had had few opportunities to use English for spoken communication, and limited exposure to English. She rated her pronunciation poorly (4 out of 10, with 1 being incomprehensible and 10 near-native), but could not specify specific sounds she had trouble with. Whilst the participant had tried shadowing before in China, she claimed she had not been able to do it well. She had only ever given one presentation in English before, and rated her level of anxiety about giving presentations in English as 7/10 (1 being very low and 10 being very high).

Analysis of the diagnostic test indicated that the participant had difficulty with two phonemes: / θ/ (i.e. voiceless *th)* and word-final /s/. In the diagnostic test and the pre-interview, there were 5 occurrences of words contains voiceless th (all of which were the word "think"), and the participant mispronounced the sound as /s/ in all instances, giving a percentage of inaccurate occurrences of 100%. In the same tests, there were 42 occurrences of word-final /s/, of which the participant pronounced 14 incorrectly, giving a percentage of inaccurate occurrences of 33.3%.

Pronunciation improvement was measured by the number of accurate/inaccurate occurrences of the voiceless TH sound and word-final S sound in the diagnostic test and pre-interview as compared to the presentation rehearsal and in-class presentation. It should be noted that the participant's summary contained 6 instances of voiceless th and 6 instances of word-final s, and the shadowing log indicated that she had shadowed each phoneme a total of 84 times.

In the presentation rehearsal, there were 22 instances of the voiceless th, and the participant pronounced the sound incorrectly only twice, with both mistakes occurring when she was visibly struggling with the presentation and new vocabulary (mistakes were made with the word "thinking" and "the"). The percentage of inaccurate occurrences was 9.0%. In addition, analysis indicated that the participant used 2 words with the voiceless th in her rehearsal that she had not shadowed. Both words (two repetitions of "thanks") were pronounced correctly, which Haufe suggests may mean that the participant carried over learning of voiceless th from shadowing.

There were also 43 instances in which a word-final /s/ should have occurred, of which the participant pronounced 34 accurately and 9 inaccurately (20% inaccuracy rate). This is a percentage of inaccurate occurrences of 20.9%. Haufe notes that inaccurate pronunciation of the sound did not appear to be related to stumbling in the presentation, but appeared random and unexplainable. In addition, analysis indicated that the participant used 6 words with the target sound that she had shadowed, 4 of which were pronounced accurately (33.3% inaccurate occurrences) and 37 that she had not shadowed, of which 30 were pronounced accurately (18.9% inaccuracy rate). Haufe suggests that this indicates that the participant could transfer pronunciation learnt through shadowing to other contexts.

In the presentation itself, there were 20 instances of words where the voiceless th should have been used, of which 10 were pronounced accurately and 10 inaccurately (50% inaccuracy rate). Haufe connects this to the participant's anxiety about her in-class presentation. All inaccurate productions of the voiceless th were found in words that the participant had shadowed.

There were 32 instances of words in which the word-final /s/ should have been pronounced, of which 24 were pronounced accurately and 8 were pronounced inaccurately (25% inaccuracy rate). Mispronunciation occurred in both shadowed words (2 out of 6 were mispronounced) and unshadowed words (6 out of 20 were mispronounced).

Results from the post interview and shadowing log indicated that the participant had shadowed the recording 14 times, over in 3 days, rather than 2 weeks. The log indicated that she shadowed twice the first day, nine times the second day, and three times the third day. In the interview, the participant stated that she considered shadowing a useful pronunciation activity, as it allowed her to compare her pronunciation with the researcher's, and to notice her mistakes. She rated shadowing highly as a pronunciation activity (7 out of 10, with 10 being very high), evaluated her pronunciation as 6 out of 10 in the in-class presentation, and rated her anxiety in the in-class presentation as 8/10 (with 10 being very anxious). She also believed her pronunciation was the same in the in-class presentation and the rehearsal.

From these results, Haufe (2013) concludes that shadowing probably did help the participant improve her pronunciation for her oral presentation, even with words she had not shadowed. This improvement is despite the fact that participant did not shadow 20 times over 2 weeks, but 14 times 3 days before the presentation. However, it appears that pronunciation improvements did not carry over from the relaxed rehearsal to the more anxiety-provoking in-class presentation, and that the activity did not help the participant with her anxiety. Based on these results, Haufe emphasises that pronunciation instruction should be more integrated in a more ongoing way into EAP programs, especially as oral presentations often have a grading component based on pronunciation. She also cautions that oral presentations may not provide a true picture of students' pronunciation abilities, due to the impact of anxiety, and, as such, pronunciation measures should also be taken in a more relaxed environment.

This study is a relatively detailed case study that does provide some evidence that shadowing can improve segmental pronunciation, and that learners find it beneficial. However, as well as the small sample size, there are a number of limitations to the study that should be taken into account. Firstly, there several issues related to results about pronunciation improvement. Most importantly, the pre-test measure of pronunciation improvement (the diagnostic test and the interview) do not appear entirely comparable to the post-test measures (rehearsal and in-class presentation) in terms of the level of formality of speech and variety of phonemes, which could have acted as confounding variables. Although the passage in the diagnostic test was selected due to its academic vocabulary and similar content to the presentation, it would have been beneficial to have a more comparable pre-test measure, such as the participant reading her presentation summary (as the author herself notes). This is a particularly important limitation as the pre-test only provided 5 instances of the voiceless TH sound, and all were in the same word: "think". As such, the pre-test does not seem to be a valid way of measuring the participants' ability to produce the phoneme as a whole, as her pronunciation of the sound could have become fossilised as /s/ in "think", but not in other words. In addition, although the author justifies her focus on segmentals, it also seems a missed opportunity for the study not to analyse improvement in suprasegmentals, as they are of great important for oral presentations.

There are also a number of issues related to the qualitative research component. Firstly, the fact that the researcher conducted both training and the post-interview could have biased results, as the student may not have wanted to disappoint her training by expressing negative views on shadowing. Secondly, the author does not report the method with which she analysed the interview data, for example using thematic analysis or grounded theory with the transcripts.

Because of this, although the study is highly relevant for this review, its contribution and trustworthiness are somewhat limited.

9)

Hori, T. (2008). *Exploring shadowing as a method of English pronunciation training.* [Unpublished master's dissertation]. Kwansei Gakuin University.

Hori (2008) conducted an experiment to examine whether shadowing enabled Japanese learners of English to improve listening ability and pronunciation, and whether the technique had an influence on working memory.

Participants were 43 male Japanese learners of English, aged 15 to 18. All were recruited from the same school, and their proficiency ranged from 100 to 160 of the TOEIC Bridge test, indicating a low-intermediate level (around B1). No participants had experience of living in English-speaking countries for any time period over a month.

Participants were divided into an experimental (n = 26) and control group (n = 17). The experimental group received 1 month of shadowing training, in which they received shadowing training twice a week, with each session lasting no more than one hour. In each training session, they shadowed two different paragraphs 15 times each, without seeing them. After the 15th shadowing, students were allowed to look at the text whilst listening to their own performance. After checking the text, they evaluated their shadowing on a scale from 1 (bad) to 7 (very good), and wrote down a few comments on their shadowing. Audio paragraphs were between 40 to 70 words long, and were sourced from English textbooks for Japanese junior and senior high schools. They were selected to be easy enough to repeat without a script and, for this reason, dialogue-style texts were excluded. The control group was provided no extra training.

Study part 1

The study is divided into two parts. The first part details the following.

To measure improvements in pronunciation and listening, a pre-test, post-test, and delayed post-test (1 month after the training) were administered. Each test contained 4 subtests: a listening test, an articulation rate test, a listening span test, and a reading aloud test. The listening test included 50 questions, adopted from the listening section of a preparation book for the TOEIC Bridge test, with the pre-test, post-test, and delayed post-test each using a different set of questions. The articulation rate test used two blocks, one of 50 monosyllabic words and the other of 50 polysyllabic words, which participants had to read aloud as quickly as possible. The same words were used for all tests, although the order was changed. In the listening span test, participants had to memorise the last word of sentences when listening to them. It was theoretically grounded in Daneman and Carpenter's (1980) reading span test, and adopted from a study by Ushiro and Sakuma (2000). Sentences in each test were different, but

were equalised for difficulty in terms of syntactic structures and vocabulary. In the reading aloud test, participants read a paragraph twice after practising for one minute. The paragraph was not one used in the shadowing training, and the same material was used for all tests. The paragraph was selected because the vocabulary and syntactic structure were sufficiently easy for the participants to read.

To analyse pronunciation, read aloud recordings were evaluated impressionistically by two native speakers of English who were experienced English teachers living in Japan. Segmentals and prosody were rated using a six-point scale. Prosodic criteria were adopted from Anderson-Hsieh, Johnson and Koehler (1992), and included stress, rhythm, intonation, phrasing, and overall prosody. Raters used a prosodic scale with 0 representing heavily accented speech that is unintelligible, 5 representing near native speech, and 2.5 representing accented but intelligible speech. Raters undertook practice rating before completing the experimental ratings. Interrater reliability was calculated, and was fair ($r = 0.48$), but meant that scores had to be averaged.

As well as impressionistic ratings, acoustic analysis was used to investigate rhythm of speech samples, that is the production of stressed vs. unstressed vowels. Pitch and duration were investigated, as they are two of the main correlates of stress in English (with intensity/volume being the third). Duration was measured in seven words selected from the reading paragraph which showed a contrast in stressed and unstressed vowels. The duration of these stressed and unstressed vowels was measured by spectrograms in for the words in the pre-test, post-test and delayed post-test, and the ratio of the unstressed to stressed vowel in a word was calculated and compared to ratios of native speakers. To measure pitch, the fundamental frequency for primary stressed vowels and unstressed vowels in five of the seven target words (two were excluded due to difficulty tracing F0 contours) was measured at the peak by using Praat. The ratio of the unstressed vowel to the primary stressed vowel was calculated to obtain the fundamental frequency, and was compared to those of native speakers. Pitch range was obtained by calculating the difference between the maximum and minimum fundamental frequency of a tone group, from 6 tone groups selected for analysis.

For the listening test results, an ANOVA analysis showed no significant main effects for group or time. For the articulation rate results for monosyllabic words, ANOVA results showed a significant main effect for time, but not for group. However, there was an interaction between time of test and group, with post-hoc comparisons suggesting a significant difference between the shadowing and control groups at post-test ($p<.05$) and delayed post-test ($p = 0.95$). In addition, for the shadowing group significant differences were found between pre- and post-tests ($p<.01$) and between pre-tests and delayed post-tests.

For articulation rate of polysyllabic words, an interaction between time and group was also observed ($p<.01$). Post-hoc comparisons suggested that the articulation rate of the shadowing group was significantly lower than that of the control group at pre-test, but increased to reach a rate significantly higher than that of the control group at post-test ($p<.05$). However, this significance did not last until the delayed post-test. Overall, articulation rates for both types of words only increased significantly in the shadowing group and not the control group.

For listening span test results, ANOVA analyses showed no significant interaction between time and group ($p>.05$). However, there was a significant main effect for time of test ($p<.05$). A post-hoc test showed that, for both groups, scores remained the same between pre- and post-tests, before significantly decreased from post-test to delayed post-test.

For read aloud results, ANOVA analyses showed that the interaction between time of test and group was not significant ($p = .139$), but that the main effect of time was significant ($p<.05$).

Acoustic analyses revealed that both shadowing and control groups produced duration differences similar to native speakers for the word "language", but that for all other words the duration ratios were higher than those of native speakers. However, the mean ratio of the shadowing group decreased from pre-test to post-test except for the word "student". The mean ratio of the control group over the same time periods, however, increased for four and decreased for three words. However, ANOVA analyses showed that there were no significant main effects nor interaction between time of test and group, implying that shadowing training did not affect durational differences between unstressed and stressed vowels.

In terms of pitch, ANOVA analyses were used to compare F0 ratios between groups over time, in comparison with native speakers F0 ratios. Native speakers produced unstressed vowels with durations of between 15% and 72% of stressed vowels. The shadowing group produced unstressed vowels between approximately 30% and 110%, whilst the control group produced unstressed vowels between 20% and 125%. No significant main effects or significant interactions were found between time and group for the words "America", "understand" or "language". However, there were significant interactions for time and group ($p<.05$) for the word "abroad", with Bonferroni tests revealing that the ratio of the shadowing group decreased slightly between pre- and post-test ($p<.05$). There was also an interaction between time and group for the word "students" ($p<.05$), with Bonferroni tests indicating a significant difference between shadowing and control groups in the post-test ($p<.05$) and results in the delayed post-test nearing significance level ($p=.059$). Overall, the shadowing group produced significantly more fundamental frequency differences in two out of five words in the post-test or delayed post-test, moving closer to the ratio of native speakers.

In terms of pitch range, ANOVA analyses showed interaction between time and group in three tone groups, indicating that changes in F0 range over the three tests tended to be different between groups. However, post-hoc tests did not reveal significant differences between levels. Despite this, in tone groups 3 and 5, the F0 range of the shadowing group showed a slight tendency to increase between pre- and post-test, whilst the F0 of the control group did not increase significantly. Overall, whilst the shadowing training seemed to have an effect on the F0 range, increasing it slightly, participants' pitch range was still much narrower than that of native speakers.

From these results, Hori (2008) concludes that, although impressionistic judgments showed no significant improvement, shadowing may have improved participants' pronunciation. She highlights that the effects of shadowing were most evident in participants' articulation rate and fundamental frequency of speech (in 2/5 target words, the fundamental frequency ratio became closer to that of native speakers, and in 3/6 intonation groups, the fundamental frequency range moved closer to native speakers). She suggests that changes in articulation rate may be due to the fact that shadowing and read aloud share most processing stages (lexical access, phonological encoding, phonetic encoding and articulation), although she acknowledges that shadowing did not increase articulation rate in a very durable way. She also hypotheses that shadowing had an impact on fundamental frequency because this was the easiest perceptual cue for the participants, as pitch is important in Japanese. She also concludes that there was no improvement in listening skills or span as a result of the shadowing training. She hypothesises that the lack of increase in listening comprehension may be due to the short timeframe of the

study, with listening improvement developing after longer periods of shadowing, and that drops in results on the listening span test may have been due to the test itself, and the fact that the sentences used for the delayed post-test were the longest and least familiar to participants.

Study part 2
The second part of the study details how shadowing recordings were analysed by the researcher.

The recordings of the 1st, 5th, 10th and 10th rounds of shadowed speech productions for texts from the 1st, 4th and 7th shadowing sessions were analysed. In total, 300 recordings were rated for analysis.

First, the author counted the number of syllables pronounced correctly in comparison with the model stimuli, and calculated the proportion of repeated syllables out of total syllables per text. A rater also rated 30 sample recordings randomly selected from the 300 recordings, and interrater reliability was high, at 0.95. Five words from a text were also analysed for duration ratio of stressed vs. unstressed syllables in the 1st, 5th, 10th, and 15th rounds of shadowed speech. Three words were also measured for fundamental frequency, with two words having been excluded due to difficulties detecting this measure. Finally, differences between maximum and minimum fundamental frequency of six tone groups were measured for the 1st, 5th, 10th and 15th shadowing.

Results suggested that the mean ratio of correctly repeated syllables per text increased the most between the 1st and 5th shadowing sessions. After the 5th session, the mean percentage increased gradually until the 10th shadowing, but did not reach 100%. ANOVA analysis of mean results for 25 participants in texts 1, 7, and 13, with text and time as within subject variables, showed a statistically significant main effect. A Bonferroni test revealed that the mean percentage of text 13 was significantly different to those of Texts 1 and 7 ($p<.01$), with the mean percentages of Texts 1 and 7 being the same. In addition, the mean percentages for the 1st, 5th, 10th and 15th rounds of shadowing are all significantly different from one another, except the 10th and 15th rounds.

As for the duration ratio, ANOVA analyses indicated no significant change for the words "enjoy" and "gestures" per round of shadowing. The word "popular" showed a slight tendency of a significant main effect ($p=.094$) and "today" and "communicate" showed significant main effects over time ($p<.05$ for both). However, the ratio of "today" significantly decreased over time, whereas that of "communicate" increased.

In terms of fundamental frequency, the ratio of the three words did not change significantly over the 15 rounds of shadowing ($p>.05$).

For F0 range, no significant effect of time was observed for tone groups 1, 5, and 6 ($p>.05$). However, a significant main effect of time was observed for tone groups 2, 3, and 4 ($p>.05$).

Hori (2008) concludes from these results that the effect of repetition on correctly reproduced syllables was the same, increasing the most between the 1st and 5th shadowing session and continuing to increase until the tenth session. She hypothesises Text 13 had more correctly reproduced syllables as it was recorded in the 7th session, when, after 6 hours of training, participants were used to shadowing. She notes that, from these results, it appears that the rhythm of shadowed speech, as measured by duration and fundamental frequency, did not

change significantly. This contradicts the results of the previous abovementioned study. In terms of F0 range, she concludes that the significant changes moving toward the model speech are in alignment with the results of the previous study.

Overall, this study is highlight relevant to this systematic review. One of its strengths is including both impressionistic and acoustic measures of pronunciation improvement, which many studies reviewed do not attempt. In addition, its definitions of shadowing and links to the theoretical literature about how shadowing may have improved pronunciation are strong, although it should be noted that they are tentative hypotheses.

However, the results from the study should not be over-interpreted. Hori (2008) claims that the study provides evidence that shadowing does improve pronunciation, but, in reality, results from the study are weak or even mixed. It is important that impressionistic ratings showed no improvement in pronunciation in terms of stress, rhythm, intonation, phrasing, and overall prosody and intelligibility. As the benefit of improving pronunciation is being able to communicate more clearly with real people in real-world situations, this result should not be taken lightly. In addition, results from acoustic analyses are mixed. It is true that articulation rate after training (although not at delayed post-test) and percentages of correctly produced syllables during training showed clear improvement. However, results related to fundamental frequency of speech are less clear, with only 2/5 words improving over time in the first study, and no words improving across rounds of shadowing in the second study. Results related to pitch range are also very tentative, with 2/6 tone groups slightly improving over time in study 1 for the shadowing group, and 3/6 tone groups improving across rounds of shadowing in study 2.

In addition, there are several aspects of the study design that limit the trustworthiness and generalisability of the results. Firstly, there is no mention of whether the groups were taught by the same teacher, which is problematic as the teacher variable is a powerful one. Secondly, there is no mention of whether all of the different kinds of human rating procedures (for impressionistic judgements or correctly pronounced syllables) were blind with regard to the time of the recording. This is problematic as it could have introduced an element of bias into the results, especially given the fact that the researcher herself was the core rater for correctly pronounced syllables. Finally, it is questionable whether ecologically valid improvements in pronunciation can be measured by read aloud tasks and shadowing itself, and for this reason it would have been advisable to include a spontaneous measure of pronunciation in the study.

10)

Horiyama, A. (2017). The development of English language skills through shadowing exercises. 文京学院大学外国語学部文京学院短期大学紀要 [Bunkyo Gakuin Faculty of Foreign Studies Junior College Bulletin], 12, 113-123.

Horiyama (2012) investigated student perceptions of a shadowing intervention to improve English skills. Participants were 25 freshman students from a university in Japan, with a low-intermediate level of English. No information is given about their L1, which is assumed to be Japanese, or their gender.

The shadowing intervention lasted 3 months, and used audio materials from the textbook *Issues for Today*. Each round of shadowing involved the same steps. After studying passages from the textbook, students undertook a shadowing activity, guided by the following protocol:

Step 1: pair-work. Students listened to the CD and analysed chunks of meaning, prosody and pronunciation. They practised pronunciation and reading with a partner.
Step 2: Shadowing with focus on prosody twice. Students shadowed the passage concentrating on enunciation and intonation, rather than meaning or content.
Step 3: shadowing with focus on meaning. Students shadowed with meaning in mind, and were encouraged to picture the story in their head and anticipate the content of the story whilst shadowing.
Step 4: shadowing with setting personal learning goals. As they had shadowed the passage four times, students were encouraged to set personal objectives based on their performance so far. They then conducted a full round of shadowing with this goal in mind.

After the 3-month intervention, the researcher conducted an eight item survey with the participants to explore their views on the effectiveness and enjoyability of shadowing. The survey questions were a mix of closed questions, open questions, and Likert scale questions.

Frequency tables were created to analyse the results. In total, 20/25 students marked that shadowing was effective or highly effective for the evaluation of prosody. 23/25 students marked that shadowing was effective or highly effective for listening comprehension. 21 out of 25 students marked that it was effective or highly effective for reproduction. 23 out of 25 students marked that shadowing was effective or highlight effective for speaking. Students also mentioned, in responses to open questions, that they thought shadowing was effective for prosody, listening, reproduction, and speaking in their comments to an open response question. 5 students did not enjoy shadowing, commenting that it was too hard, and that the materials were too fast.

Based on these results, Horiyama (2017) concludes that students enjoyed shadowing and considered it effective. She also presents several hypotheses connected to the results. Firstly, she believes that that shadowing can help student motivation by developing communicative skills and helping students see progress in their English skills. Secondly, she claims that shadowing can create an active learning environment and that setting learning goals as part of shadowing is beneficial for students. Finally, she highlights that instructors should use materials tailored to the students' level to promote confidence and improve skills.

It seems valid to conclude that students enjoyed the intervention and found it effective, based on results. It also seems reasonably to conclude that setting learning goals was beneficial for students, as, although students were not asked directly about this in the survey, setting personal goals was a key part of the shadowing procedure. Based on the negative comments from 5 students about the difficulty of the materials and speed, advising teachers to assign materials according to student proficiency also seems in line with the findings. However, Horiyama's connection between shadowing and motivation is not directly in line with any of the findings of the study, so should be taken as a personal belief.

Key issues limiting the generalisability of these results are the small sample size, lack of investigation of the internal consistency of the survey, and the very limited analyses conducted. The only measure used to analyse the results were frequency tables, with no means or standard deviations presented, and no statistical analyses performed to assess the significance of the

results. In addition, measuring student perceptions through a questionnaire provides limited information, and richer, more granular detail could have been provided though semi-structured interviews.

The fact that the survey may have been conducted by the teacher themselves may also inflate positive results, as students may have been eager to please their teacher or afraid of negative consequences related to critical responses.

11)

Hsieh, K., Dong, D., & Wang, Y. (2013). A preliminary study of applying shadowing technique to English intonation instruction. *Taiwan Journal of Linguistics*, 11.2, 43-66.

Hsieh et al. (2013) explored the effect of shadowing training and practice on students' scores in fluency, word pronunciation, and intonation.

Participants were 14 non-English major students at Taiwan National University, who were randomly and equally assigned to a control and experimental group. The computer program My English Teacher (MyET) was used to measure fluency, word pronunciation, and intonation as a pre-test, with no significant differences detected between the groups. It is unclear what this pre-test involved, and whether it collected controlled or spontaneous speaking samples. After the pre-test, the experimental group received an 8-hour training program in shadowing. It is unclear if this training was provided weekly, monthly, or at another interval. The control group, however, received no training program and practised lessons from MyET following standard repetition techniques. It is not stated clearly whether these lessons are the same as those used in the shadowing training in the experimental group, and whether the total duration of practise for both groups was the same. At the end of the semester, MyET was used for a second time to take the same measures of pronunciation described above.

Means and standard deviations were calculated for pre- and post-test results, and suggested that the experimental group outperformed the control group by around 8-9 points, both overall and in terms of fluency, word pronunciation, and intonation. An independent samples t-test confirmed that these differences were all significant.

The authors conclude from these results that shadowing is effective for training pronunciation, and should be considered as an alternative to the widespread method of repetition in Taiwan. They hypothesise that shadowing may be more effective than repetition as familiarisation with texts in shadowing reduces demands on short-term memory.

Whilst the authors admit that the study is exploratory in nature, there are several reasons why the evidence it presents should be taken with caution. First, description of the exact shadowing intervention followed, both in training and further practice, is lacking. This is problematic as the authors do not clearly state whether the control and experimental groups worked with the same and number of audio texts, which could influence results. In addition, the duration of the intervention is unclear, with reference only made to post-tests being conducted "at the end of the semester" or "two months before the end of the semester". Secondly, no justification is given for why MyET was used to measure pronunciation, or exactly how measurement and scores are calculated on the program. This means that, whilst the tool is assumed to be reliable due to the fact that it is a computer program, the validity of this measurement tool difficult to

determine. In addition, measurements of Finally, as the authors acknowledge, the small sample size also limits the generalisability of the results.

It should also be noted that no information is given directly about participants' proficiency, it is only noted that texts for intermediate to high intermediate learners were used.

12)

Huang, Y. (2018). ***The influence of speech shadowing on English word-initial consonants produced by speakers of English as a foreign language.*** **[Unpublished master's dissertation]. California State University.**

Huang (2018) conducted a study to examine the effect of a 4-week shadowing program on participants' production of English word initial consonants /b/, /p/, /t/, /d/, and /k/.

Participants were 20 children studying at an afterschool learning centre in Taipei City, Taiwan. All were L1 speakers of Taiwanese Mandarin, aged between 8 and 12 years old. There were 10 boys and 10 girls in the study. All had been learning English from the ages of 5-8, and all participants were familiar with 500 to 700 English words.

Participants were randomly divided into a control group (n = 10) and a treatment group (n = 10), and assigned an identification number based on their group membership. It should be noted that, although the amount of English use per week appeared the same in both groups, participants in the experimental group were, in general, older than those in the control group, and had therefore received more years of English education. The gender balance was also different in the two groups, with 7 males and 3 females in the control group, and 7 females and 3 males in the experimental group. Participants from both groups continued to participate in their regular English classes (2 hours, twice per week). However, whilst the control group received no speech shadow training, the treatment group received shadowing instructions and training 4 days per week for 4 weeks.

The shadowing took place during the 1-hour break period between classes, and shadowing practice was around 5-8 minutes. Before the first shadowing practice, the researcher gave a 30-minute shadowing lecture to explain and demonstrate the technique, and monitor students to ensure they were shadowing correctly. After the first session, participants practised in a quiet one-to-one setting with the researcher. Participants kept a log of how they practised and what reading sections they used throughout the study, and none shadowed at home.

Shadowing materials were sourced from the publication *Let's Talk in English*, as this publication was considered suitable for learners with around 600 words, and the soundtracks could be adjusted for speed. The content of the audio texts was related to topics young students are familiar with, such as travel, food, and landscapes. The texts were separated into 1–2-minute segments for shadowing. Texts were played at 0.85 in the first shadowing practice, 0.95 in the second, and then at full speed. Participants shadowed on laptops with earphones for with 3-4 soundtracks in each session.

Pre- and post-tests were designed to collect samples of participants' speaking. The pre-test involved students reading a passage out loud and then answering questions about it orally. Questions were three fill-in-the-blanks. The passage was designed following sentences structures and vocabulary from *The Plan for Implementing English Language Teaching in Taipei City Elementary Schools: Curriculum, Guidelines of Elementary English Language*

Teaching and Learning (Department of Education, Taipei City). The question-answer format was deemed most suitable to collect speech samples as it was difficult to get the students to discuss the reading passages more freely. The post-test was different to the pre-test, as some students memorised the passage from pre-test. In addition, the students' teacher explained that the level of English in the pre-test passage was too low for students. For this reason, the post-test passage was designed to be more difficult, on the topic of dinner. However, the format of reading out loud then responding orally to fill-in-the-blank questions was the same. In addition to the reading and speaking tasks, students in the experimental group were also recorded doing shadowing at post-test. The words used to calculate Voice Time Onset largely followed CVC or CVV patterns, as the teacher informed the researchers that student had difficulties with CCV patterns.

Recorded data was analysed with Praat software, to calculate values of VOT for the five consonants under study. VOT values were then averaged, to reduce the effect of the vowels following them on their production, and compared between pre- and post-test. VOT values of American English word-initial stops from Lisker and Abramson's (1964) research were used as a standard of comparison, as there were no opportunities to recruit native English speakers of the same age as participants for comparison. VOTs from the experimental and control groups' reading and speaking performances at pre- and post-test were compared with a Mann-Whitney U test. Differences between the experimental group's pre- and post-test were analysed descriptively, due to the fact that the post-test contained a component (shadowing) not present in the pre-test, and the fact that participants may have been fatigued due to the more difficult reading passages and extra task.

Results indicated that, at pre-test, there were no significant differences between the control and experimental groups in the pronunciation of /b/, /p/, /t/, /d/ or /k/ in the reading condition. In the speaking condition, the control group pronounced /t/ significantly closer to standard values that the experimental group. Pre-test results also showed that few participants in the reading condition, and no participants in the speaking condition, produced /t/ or /d/ within the standard average for native speakers (five participants for /t/, and one participant for /d/ in the reading condition). When comparing averaged VOT values of all consonants to standards, only the control group's average production of /t/ reached the standard value.

At post-test, the control group, on average, pronounced /b/, /t/, /d/, and /k/ closer to average VOT values than the experimental group, in both speaking and reading conditions. The experimental group pronounced /p/ as closer to average VOT values than the control group. In addition, only one participant, from the control group (A2), pronounced a consonant at a standard value (/t/ in the reading condition).

At post-test, the experimental group's production of /p/ significantly improved, moving towards native-like values, with a large effect size ($r = 0.82$). There were no other significant differences between the groups. In the speaking condition, the experimental group showed significantly more improvement in their production of /b/, /p/ and /d/ than the control group. However, only the effect size value of /p/ (0.98) reached the recommended level of 0.8. In addition, despite the experimental group's improvement, the control group produced /b/ and /d/ as closer to standard averages.

For the experimental group, VOT values at post-test were different across conditions (reading, speaking, shadowing) for all consonants but /d/. VOT values for four consonants (/b/, /p/, /t/, /k/) were closer to standard averages in the shadowing condition than in the reading and

speaking conditions. /b/ and /k/ were closer to standard values in the speaking condition than the reading condition, and /p/, /t/ and /d/ were closer in the reading than speaking conditions.

Results from the post-test questionnaire with the experimental group indicated that five participants found shadowing very interesting, three thought it was interesting, one found it moderately interesting, and one thought it was boring. Nine participants stated that shadowing was difficult, whilst one thought it was easy. Two participants felt their comprehension of texts ranged from 80% to 100%, four between 60-80%, three between 40-60% and one between 20-40%. Six participants believed they had repeated every word similarly to what they had heard, whilst four disagreed with this statement. Three participants believed their listening skills improved significantly, six that they improved a little, and one that their skills did not improve at all. Eight participants believed their vocabulary improved, seven that their pronunciation improved, four that their grammar improved, and seven that they learnt about sentence structured. One participant stated that the tracks were too fast.

The author concludes that these results provide some evidence that shadowing may assist learners in improving their pronunciation of consonants. She also suggests that it may help with other areas of proficiency, like vocabulary, based on the survey results. However, she cautions that more research is needed into the latter area of research.

The author does acknowledge several limitations of the study. For example, the effect of the small sample size on the results of post hoc power analyses and effect sizes and how this indicates larger samples are needed. She also highlights the short duration of the study and the fact that participants' proficiency should have been controlled for more closely.

However, there are a number of limitations that she does not acknowledge. Firstly, there are several issues related to study design. For instance, there appear to have been significant differences in the control and experimental group in terms of age and years learning English, which were not controlled for and could have impacted the results. It should also be noted that the experimental group received shadowing in addition to normal classes, in a study break with the researcher, and the control group had no time with the researcher and no extra activity in place of shadowing. It would have been more comparable for the intervention to have taken place in normal class time. In addition, the pre- and post-tests were significantly different to each other in terms of difficulty, which could have influenced results. Whilst the researcher changed the post-test to adapt to students' level, this should have been explored prior to the study with, for instance, piloting of the pre- and post-tests. Ideally, these tests should have been comparable in terms of language and topic difficulty, and counterbalanced as to when each was presented to students. It would also have been beneficial to include more spontaneous tasks suitable for the learners' age and proficiency to collect pronunciation samples, for example picture description tasks. Additional measures of pronunciation would have also been beneficial. In terms of the questionnaire, it should be noted that no information is provided on its design, and it appears to have been conducted by the researcher. This latter fact could have biased results, as it could have meant students presented overly positive views in order to please their teacher.

At a more theoretical level, there is also no explicit connection made between how improvement in the consonants studied could improve participants' overall pronunciation or measures of comprehensibility or accentedness. Ecological validity of results is therefore limited.

For these reasons, although the study does have a number of strengths, such as the detailed description of the shadowing intervention followed, its findings should be taken as tentative, and mixed.

13)

Huang, H., Barrett, N., Lo, M., & Tseng, C. (2023). The effectiveness of shadowing practice with web-based apps: towards promoting the comprehensibility and lexical-level intelligibility of EFL students' presentations. ***English Teaching and Learning.*** https://doi.org/10.1007/s42321-023-00145-w

Huang et al. (2023) conducted a mixed-method quasi-experimental study to investigate the effect of web-based shadowing practice on participants' comprehensibility and intelligibility when giving presentations in English.

The study involved 90 participants from a university in southern Taiwan. All were college juniors, with an L1 of Mandarin and an English level ranging from A2-B1 level.

Three intact classes, from a course on professional English for communication and presentation, were used. Two classes were used as experimental groups, and one group was randomly selected as the control group. All groups received the same course over the semester, with the exception of a 6-week shadowing intervention, which was received only by the experimental groups.

The intervention had two components: in class pronunciation instruction and shadowing practice with Google Translate and Oddcast. Google Translate was used to practice pronunciation of single words, and Oddcast was used to practice pronunciation of full sentences. Materials for shadowing were selected from participants' own presentation scripts, which they wrote as part of their professional English course. In week one, participants read their presentations out loud. Instructors marked their presentations and selected five sentences with the highest number of incorrectly pronounced words. These sentences were used as both shadowing material and as pre- and post-tests in week one and after week six, respectively.

Four expert raters listened to the pre- and post-test recordings, and rated them for comprehensibility and intelligibility. It is unclear if they listened in a randomised order. Comprehensibility was measured on a 9-point Likert scale (1= extremely easy to understand, 9 = impossible to understand). Lexical-level intelligibility was measured by raters underlining mispronounced words on a script, then dividing the number of correctly pronounced words in the recording by the total number of words. Interrater reliability was calculated, and was high for both measures.

To analyse the results from pre- and post-tests, a Mann-Whitney U test was applied. This was because the scores of both comprehensibility and lexical level intelligibility were considered ranking and had non- normal distribution. No differences were found between groups in comprehensibility at pre-test. However, post-test scores for experimental group were significantly different to control group (p = .001 <0.5), indicating that shadowing practice contributed to significant improvements in comprehensibility compared with conventional instruction. In terms of lexical-level intelligibility, no differences were found between groups at pre-test. However, post-test scores for experimental group were significantly different to

control group (p=.002 <0.5). This indicates that instruction with shadowing practice enhanced lexical-level intelligibility more than conventional instruction.

In addition to pre- and post-test scores, the study also involved a 6-item questionnaire with 58 students. The questionnaire showed high internal consistency, with a Cronbach's Alpha score of .84. Means and percentages were calculated for each of the items, although it should be noted that the scales used for each question are not described (from the results they could be assumed to be Likert scale questions in which 5 is strongly agree). Results showed that all students strongly agreed that shadowing practice was useful and helpful (mean 4.16), and that recording speaking helps to improve speaking and pronunciation (4.14). They also agreed that the shadowing intervention helped to improve pronunciation (4.09). Finally, Google Translate and Oddcast were viewed positively for helping improve pronunciation and intonation. In qualitative analysis of the two open ended questions, four recurring themes emerged: (1) effectiveness of instruction, (2) use of metacognitive strategies i.e. students noticing pronunciation errors and making improvements, (3) motivation and autonomous learning e.g. the intervention motivated them to study more, (4) time and effort commitment, with some students complaining the assignment was too time-consuming and two saying it was tedious.

Finally, semi-structured interviews were also carried out with 18 randomly selected participants from the experimental group. Four themes emerged in the analysis of the interview data: (1) innovation and usefulness of technology tools, (2) time spent on assignment (some mentioned practice time was too long, but most said they appreciated being pushed to practice more), (3) awareness-raising in pronunciation, (4) pronunciation improvements after training (most students said that their pronunciation had improved in word stress, intonation, and fluency).

Overall, the pre- and post-test scores provide convincing evidence that the shadowing intervention helped improve students' comprehensibility and intelligibility in their presentations. The only issue, related to the generalisability of the results to overall comprehensibility and intelligibility, is the repetition effect that may be present in using the same sentences as pre- and post-tests and shadowing practice. This raises the question of whether students' overall intelligibility and comprehensibility truly improved, or just their ability to reproduce the five given sentences. The ecological validity of such a controlled speed sample is also questionable.

The data presented from qualitative findings also provides support for implementing shadowing: students appeared to enjoy the intervention and consider it effective, despite noting that it was time-consuming. Students also believed it gave them tools to monitor and improve their own pronunciation. However, these findings could be strengthened by a more rigorous discussion of the methods used to analyse the qualitative data, which is very limited. Analysis of open responses is simply said to be based on "recurring themes", and the type of analysis used for interview data is not mentioned at all.

14)

Hutchinson, A. (2022). Individual variability and the effect of personality on non-native speech shadowing. *JASA Express Lett, 2*(6). https://doi.org/10.1121/10.0011753

Hutchinson (2022) explored the effect of the Big Five personality traits on participants' imitation of the French front rounded vowel sounds /y/ and /u/.

Participants were 75 university students (15 males and 57 females, aged between 18-33) who were native speakers of American English and had had no exposure to or instruction in the French language. Participants self-reported an intermediate level of proficiency in other languages (3.22 on a 7-point Likert scale).

Participants completed the Big Five personality inventory survey, and an imitation task in which they repeated 26 French words after hearing them. 12 of the words contained the target /y/ and /u/ sounds, and the others were filler words.

Participants' imitation was recorded and compared with recordings of the same words by a native French speaker. Recordings were manually annotated in PRAAT. F1 and F2 values at the mid-point of each vowel were extracted using an LPC-based algorithm, then transformed to Bark using the PHONR package in R. After extraction and normalisation, F1 and F2 Bark values of each participant were used to calculate the degree of similarity to the model talker's production of the same item, with the participant's production values subtracted from the model talker's value of the same item. In order to explore the effects of personality traits on vowel production, four mixed effect models were performed using the lme4 package in R.

The mixed effect models showed that, for F1 values in production of /y/, there were significant effects of extraversion ($p = 0.05$) and neuroticism ($p = 0.03$). As participants' scores in these traits increased, the distance between their production and the model production decreased. However, no correlation was found between any of the Big Five traits and F2 values in productions of /y/. For production of /u/, there were significant correlations between extraversion ($p = 0.03$) and neuroticism ($p = 0.03$) and F1 values. No correlation was found for any of the Big Five traits and F2 values. Hutchinson concludes that, based on her results, shadowing of /y/ and /u/ was significantly influenced by the personality traits of extraversion and neuroticism.

Whilst the data collection in the study appeared reliable, due to the standardised procedure and comparison with the same model speaker, and valid, due to the use of a validated survey (the Big Five personality inventory), the study's results are not relevant to this systematic review. This is because the task carried out does not meet the core principles of shadowing as defined in this review: repeating what is heard as simultaneously as possible, and speaking whilst listening to incoming information. The task that participants carried out was, essentially, a classic imitation task. For this reason, Hutchinson's (2022) study is not discussed in the results section.

Another issue not addressed in the survey is the unbalanced sample, with 15 males and 57 females. Gender, personality, and culture may be linked in a number of ways, but the authors do not reflect on the fact that their sample is largely female American population and how this may have influenced their results.

15)

Kunihara, T., Zhu, C., Minematsu, N., & Nakanishi, N. (2022). Gradual improvements observed in learners' perception and production of L2 sounds through continuing shadowing practice. *Interspeech 2022*, 1303-1307.

Kunihara et al. (2022) conducted a study to explore the effects of a 42-day self-study "shadowing marathon" on participants' perception and production of English sounds.

Participants were Japanese learners of English with A1-A2 proficiency. All were freshmen/sophomore students who were majoring in Global Communication. Originally, 35 participants signed up to participate in the shadowing marathon, which was offered as an optional event over the university summer holiday period. However, 10 participants did not attend every day, leaving 20 in the study.

The marathon involved shadowing 4 oral texts every day, for a total of 168 texts over the whole study. The texts were around 30 seconds long and were selected from the EIKEN listening tests (made for learners A2-B1) to ensure similar difficulty of the passages. In addition, the same passages were used for day 1, 23, and 42, to control for the effect of text difficulty on results. For each passage, students did three rounds of normal shadowing and one round of script-shadowing. After script-shadowing, they read the passage aloud without listening to it. For the first two weeks, speed of the passages was reduced to 0.8, for the next two 0.9, and for the last two the original passages were used. Participants' shadowing, script-shadowing, and reading were recorded.

These recordings were subjected to acoustic analysis to explore to what degree participants' productions were in line with the model recordings, in terms of segmentals and prosody (measured by duration, pitch, intensity). Analyses were also used to explore improvements in listening perception skills, which were explored via the same recordings, with calculations of average shadowing and script-shadowing scores on Phonetic PosteriorGram-based Dynamic Time Warping (PPG-DTW). Measurements were analysed for improvement in segments and prosody within and over sessions. Statistical analyses were then used to detect any significant changes (e.g. two-way ANOVA post-hoc multiple comparisons).

Results indicated that shadowing was effective in improving perception skills. Segmental scores for in-session standard shadowing became significantly closer to the scores for script-shadowing every day, with more rounds of shadowing ($p<.001$). This indicated that, by shadowing repeatedly, listening breakdown was reduced. Across session improvements were clear, with shadowing productions becoming especially close to script-shadowing productions on days 23 and 42. Prosody in standard shadowing also generally became closer to that of script-shadowing within sessions and across sessions ($p<.05$ for both, respectively), with the exception of pitch measurements, which showed more mixed results.

Results for improvement in production were more mixed and inconclusive. Within sessions, that is, between rounds of shadowing, learners' articulatory control for segments become significantly closer to the model control ($p<.01$), except for day 23. However, segmental control for script-shadowing and first readings aloud were significantly closer to the model ($p<.001$) than standard shadowing. This indicates that learners' segmental articulatory control in session relied heavily on the written text. Comparing the first round of shadowing to the model on day 1 and day 23, and day 20 with day 42, articulatory control became significantly closer to the model ($p<.001$). However, when comparing script-shadowing and first reading on day 1 and 21, and day 20 and 40, no significant differences were observed. Comparing in-session measurements (that is, comparing the first round of shadowing to the third round of shadowing) of intensity, pitch, and duration, no significant improvements were observed. Even when text cues were available, with script-shadowing, significant improvements were only

made in a few cases. In addition, in almost all cases scores tended to move significantly away from the model between script-shadowing and first reading (p<.001). This indicated that, it was difficult for learners to duplicate the prosodic control of the model, and that pitch control was very difficult for the learners. In terms of changes across sessions, there were significant improvements in all three prosodic features in standard shadowing from day 1 to day 23 (p<.05), but no significant differences from day 20 to 42. In addition, in script-shadowing and first reading, significant improvements were observed in pitch (p<.05) but only from day 1 to day 23.

From these results, the authors conclude that shadowing was effective for perception skills, but not production skills, because of lack of explicit instruction. This conclusion seems overly dismissive of the positive results of the study, which should be highlighted: learners made improvements in segmental control over the duration of the study in standard shadowing, and in prosodic control from day 1 to 23. Whilst these are far from enough to provide strong support for shadowing as a tool to improve pronunciation, the mixed results from this study do indicate that the relationship between shadowing and pronunciation improvement may be complex. It also seems unwarranted to claim unequivocally that the mixed results were due to lack of explicit instruction, which is beyond the scope of the data presented. This certainly does seem like the most likely hypothesis, but should be represented as such, rather than a clear conclusion.

The mixed results of the study should also be taken in combination with the fact that the study did not include a control group, which would have allowed for more contextualised interpretation of the results. In addition, there are several limitations to the study that should be acknowledged. Firstly, it is noted that all students had studied shadowing before, but no consideration is given as to how this may have affected results. In addition, the issue of 15 students dropping out, and the potential self-selection bias this may have created, is left unexplored. Finally, it may have been useful to include other data collection instruments for collecting more realistic pronunciation samples for more ecologically valid conclusions, such as spontaneous speaking tasks. Other measures of pronunciation improvement, such as impressionist ratings, could have also been useful to add more depth and real-life applicability to the findings.

16)

Kuo, Y., & Chou, T. (2014). Effects of text shadowing on Taiwanese EFL children's pronunciation. *Asian EFL Journal, 16*(2), 11-43.

Kuo and Chou (2014) investigated the effects of a text-based shadowing program on students' pronunciation at word-level, sentence-level, and overall. They also explored the effect of shadowing at different proficiency levels (lower, intermediate, and higher) and student attitudes towards the intervention.

The researchers designed a mixed-method study, involving pre- and post-tests to measure pronunciation improvement, and teacher observations and questionnaires to explore student attitudes.

Convenience sampling was used to recruit four intact classes of fourth graders in a public elementary school in New Taipei City, Taiwan. Classes were randomly and equally assigned to experimental or control groups. However, one control group was later excluded, due to the

fact that it contained a large number of students with learning difficulties. This made the number of students in the control and experimental groups unbalanced (n = 26, and n = 53, respectively).

The experimental group received a 12-week intervention in which they practised shadowing for 10 minutes per day, four days per week. Materials used for shadowing included normal speech (125-150 words per minute), chants, drama scripts, and excerpts from the film *The Lion King*. All classes were taught by one of the researchers, who followed a procedure in which shadowing texts were first introduced and explained each week, before students began to practise shadowing together and individually. Students also had opportunities to perform their shadowed texts in groups in front of the class. The control group received no shadowing training, and instead did English homework.

Improvement in pronunciation was measured by a 100-word Reading Aloud Test developed by the researchers. The tense contained 100 words in 28 simple sentences, with 88 words from a third-grade textbook and 12 from the current textbook students were using. The authors justify the use of such a test with reference to similar studies using the same test format, and piloted and revised the test to develop its validity. Students' reading of the test was recorded, and scored by two different ESL experts, with scores given for correct pronunciation of words (including correct pronunciation of phonemes and stress) and sentences (including correct pronunciation of intonation and chunking). Interrater reliability was calculated, and was very high at 0.99 at word level and 0.98 at sentence level.

Results showed that there were no significant differences between groups at pre-test ($p > .05$). However, the experimental group significantly outperformed the control group at post-test on measures of word, sentence, and overall pronunciation ($p<.05$). In addition, gain scores showed significant intergroup differences in pronunciation at word level, sentence level, and overall, suggesting that the experimental group made significantly more improvement than the control group.

All proficiency levels in the experimental group scored significantly higher than the control group on post-test pronunciation at word, sentence, and overall levels. Gains were significant among all proficiency levels in the experimental group, although intermediate students improved significantly more than lower-level students, who themselves improved significantly more than higher level students. However, anecdotal observations from the researchers grading the recordings suggested that high- and low-level students in the experimental group improved in ways not encompassed by the grading criteria, for example courage to speak English (lower level) and native-like accents (higher level). In comparison with the control group, lower and intermediate level students in the experimental group showed significantly higher gains at word, sentence, and overall level, whereas higher level students did not.

Teacher observations of student achievement and attitudes, recorded once per week, were used to give context to these results. These observations suggested that shadowing was so helpful as it helped improve oral fluency in a short time (students adjusted to the fast speed of the recordings after just 4 weeks) and helped students make sound-word connections. In addition, the teacher tentatively hypothesised that enjoyment of the activity was correlated with improvement, as four students observed to most enjoy the activity were also those who made most progress post-intervention.

Results from the survey suggested that participants had positive attitudes to shadowing (77% thought they made significant progress in English pronunciation, 82.7% viewed it as useful way to improve English pronunciation, 73.1% were willing to use it at home to practice English pronunciation). The survey results also indicated that chants were the most preferred materials (62% liked chants the most, 25.9% liked drama scripts, and 7.4% movie scripts the most). Teacher observations confirmed this, showing that students got used to chanting very quickly, and enjoyed it a lot. They also seemed to be excited about learning drama scripts and movie scripts. However, the teacher observed negative attitudes and complaints about the film, suggesting the material was too difficult for the students.

Overall, the study is rigorous in its design and implementation, and contributes robust evidence to suggest that shadowing can be an effective means of improving pronunciation amongst young learners. Amongst its key strengths is its detailed description of the shadowing materials and intervention followed, with each day's work outlined clearly, and its exploration of the effects of the training for different proficiency levels. However, there are some limitations of the study that must be highlighted. Firstly, as the authors note, the number of participants assigned to the control and experimental groups was unbalanced, which could have impacted results. Secondly, as the teacher was the one who conducted questionnaires with the students, results could be overly positive, as students could have sought to please their teacher with their responses. Finally, claims that shadowing increases courage to speak amongst lower-level students, and native-like accents amongst higher students, should be taken with caution. These claims stem from comments made by researchers grading students' recordings, and were not explored in a systematic way either through the research design or in an ad hoc manner. As such, they should therefore be treated as very tentative hypotheses to be explored in future work, rather than key outputs of the research. This lack of systematic examination of qualitative evidence is also present in the teacher observations, as the authors provide no explanation of how these observations were analysed. Finally, use of a task to collect spontaneous speaking would have improved the ecological validity of results.

17)

Kurniawan, H., Sitohang, B., & Rukmono, S. (2019). Gamification of mobile-based Japanese language shadowing. ***Proceedings of the 2019 International Conference of Artificial Intelligence and Information Technology*****, 215-219.**

Kurniawan et al. (2019) conducted a study to explore how gamification of a shadowing app could improve student motivation when learning Japanese.

40 participants took part in the study, and were all aged between 20-30 years old. No information is given about participants' L1, gender, educational level, or the country in which they were recruited. All participants had an interest in Japan, but all were complete beginners in Japanese. Participants were divided into a control (n = 15) and experimental group (n = 25).

The experimental group completed shadowing exercises using an app, Shadowingu, which included 16 gamification elements, such as storytelling, experience points, rewards, badges, a leader board, and avatar. The control group completed the same exercises using a version of the app, Shadowingu Lite, without the gamification elements. It is unclear how many shadowing exercises participants completed, and over what time period. However, the learning material is described as seven lessons with 10 target words per lesson. It is also unclear what shadowing involved. That is, if it involved repetition of single words, or if words were

presented in sentences. This is important for this review as, if the words were presented alone, the study may not fit the definition of shadowing used.

To collect data, three different instruments were used: pre- and post-tests in which participants spoke a few words before and after shadowing; a questionnaire comparing Shadowingu and Shadowingu Lite, and an Instructional Materials Motivation Survey. It should be noted that information on the pronunciation tests is not given, and it is simply specified that participants had to "speak a few words". There is no reference as to how words were selected, whether pre- and posts-tests were the same or different (and if different, how they were counterbalanced for difficulty), and how pronunciation was rated. Information on the two questionnaires is also minimal, with no annexes provided and no information given as to how the comparison questionnaire was developed and tested for validity and reliability. The Instructional Materials Motivation Survey appears to have been taken from another study, but again validity and reliability is not addressed directly.

In terms of data analysis, as mentioned, no information is given on how pronunciation data was analysed. Percentages are given for whether participants preferred Shadowingu or Shadowingu lite, with open response questions from the preference survey categorised qualitatively. An independent samples t-test was conducted on data from the Instructional Materials Motivation Survey to compare the results of the control and experimental groups, with Cohen's d used to assess effect size.

Results from pre- and post-test suggest that 60% of word pronunciation improved after shadowing practice, although it is unclear both if this figure refers to the control, experimental group, or both. From the preference survey, results indicated that 92% of respondents preferred Shadowingu app, with reasons given related to direct feedback (n = 8), being more challenging (n = 4), and being more interesting/fun/motivational (n = 19). In addition, all 16 gamification elements that were applied appeared to have a positive effect on overall learning experience, although three (rules, story, avatar) were rated 3 and 4 on the Likert scale (neutral, positive). Finally, results from the Instructional Materials Motivation Survey indicated that there was a significant difference in the control and experimental groups in terms of attention, relevance, and satisfaction in the Instructional Materials survey, with a medium to large effect for "attention" and "relevance" and a very large effect for "satisfaction".

From these results, the authors conclude that shadowing is effective for practising speaking skills, and that gamification elements increase students' motivation and make the learning experience more interesting. The first of these conclusions cannot be considered trustworthy, given the lack of information about what the pre- and post-tests involved, how they were developed, and how they were rated. The other conclusions do seem warranted from the data presented, with participants appearing to show a clear preference for the app involving gamification, and the experimental group showing significantly higher attention and satisfaction with the app, as well as feeling it was more relevant. However, they should also be taken with caution due to the small sample size in the study, the unbalanced control and experimental group, and the lack of information given about the participants and the time period of the study. Importantly, if shadowing tasks involved simply listen and repeat of individual words, the study is not relevant to this systematic review.

18)

Lin, L. (2009). ***A study of using "shadowing" as a task in junior high school EFL program in Taiwan*****. [Unpublished master's dissertation]. Taiwan University of Science and Technology.**

Lin (2009) conducted a study to explore whether shadowing could help improve junior high school students' listening and speaking performance, as well as their perception of the training intervention.

Participants were 25 students at a junior high school in Taiwan. 10 were boys, and 15 were girls, and all were aged around 14. All were native Taiwanese, although the researcher does not note if their dominant languages were Mandarin or Taiwanese. All students had received at least 5 years of formal EFL instruction in school, and none had spent more than one month travelling or studying in English-speaking countries. According to the level of the pre- and post-tests used for speaking and listening ability, participants had a basic level of English proficiency. None had received shadowing training, or practised shadowing before. Lin acted as both teacher and researcher in the study, and had been both the students' English and homeroom teacher for 1 year before the study began.

The study involved a 5-week program in which there were three 55 minute classes per week, for a total of 15 classes. Classes 1, 14, and 15 were used for testing or interviewing, meaning that there were only a total of 12 hours of actual shadowing practice. In these 12 hours, participants were trained to shadow, first in Chinese then in English, and then practised shadowing and received feedback. Participants practised with two main texts, which were sourced from the textbook *Go Super-Teens (Book II),* a resource of a similar difficulty of the participants' textbook from the year. The texts were between 100-110 words long, and between 48 seconds and 1 minute 12. The study was conducted during the summer vacation, so that participants could focus more on the task, and interference from other school activities and subjects could be minimised.

Data collection instruments included a pre- and post-test for listening and speaking, a pre- and post-questionnaire, an interview schedule, and field notes.

The pre- and post-tests were the speaking and listening sections of General English Proficiency Test (GEPT) mock tests adopted from the Language Training and Testing Centre (LTTC). This is a standardised test with high reliability, and is widely used in government institutions and private enterprises in Taiwan. As some of the questions on the pre- and post-tests were the same, participants were not provided with their pre-test grades to avoid the influence of memory effect. To evaluate speaking, participants answered 16 questions, which were a mix of repetition, reading aloud short sentences and a short paragraph, and answering short open-ended questions.

The pre-questionnaire, developed based on the literature and discussions with experts, explored participants' attitudes to learning English through 12 Likert-scale questions (presented on a 5-point scale from "strongly agree" to "strongly disagree", and one open-ended question. The post-questionnaire included two extra Likert-scale questions to explore difficulties participants faced when shadowing, and the extent to which the researcher's assistance in shadowing was helpful. Both questionnaires were administered in Chinese.

The interview guide was semi-structured, and contained five set questions, with additional questions sometimes added based on participants' responses. Questions explored reactions to

shadowing, ideas about shadowing vs. repetition, the perceived effects of shadowing, and intentions to continue to shadow in the future. Interviews were held in Chinese, and each interview session lasted around 4-5 minutes, with one interviewee in an empty classroom. Interviews were recorded and transcribed.

Field notes included observations of participants' performance and reactions to shadowing, interactions in class, and the researcher's thoughts and reflections. Video recordings were also used for more in-depth analysis of interaction in the classroom.

To analyse results, SPSS was used to analyse statistical data for scores of pre- and post-test and both questionnaires. Paired-samples t-tests were used to explore any significant differences in pre- or post-test results, and between questionnaire answers. Data from field notes, interviews and open-ended questions were transcribed, coded and categorised, following Strauss and Corbin's (1997) guidelines on reorganisation.

Results from the speaking test indicated a significant improvement (p =.000) from pre- to post-test, with mean scores increasing from 40.6 to 62.8.

Results from the questionnaire indicated no significant increases in interest in English, being active in learning English, or the value given to practising listening or speaking (which were both important to the participants even at pre-test). However, there did seem to be significant increases in participants' confidence in their listening and speaking ability at post-test, how much they imitated stress and intonation when speaking after listening, and their feeling that they could speak English actively (all p = .000). All of these claims are also evidenced with excerpts from interviews, with participants reporting, for instance, a fear of speaking that lessened after shadowing, and a blossoming confidence. In addition, participants' interest in different methods to learn English, and in shadowing, significantly increased at post-test (p = .000). Finally, survey results indicated that 28% of the participants had difficulty shadowing (agreed or strongly agreed), which interview results suggested could be due to difficulties following the oral texts and the speeds of the texts. However, 88% of students indicated (agreed of strongly agreed) that their teacher helped them overcome such difficulties, a finding triangulated by interview data.

From the qualitative data, several categories emerged to explain students' reactions to shadowing. Positive reactions were based around language ability enhancement (including improvement in listening and improvement in speaking, in terms of fluency and pronunciation), effective learning approach (including enhanced motivation and confidence), and student-centred learning (including self-learning, self-correction). Negative reactions were based around concerns that the activity was repetitive and time-consuming, less challenging, or pressurised due to recordings, and demotivation due to failure.

From these results, Lin (2009) concludes that shadowing can benefit students in EFL classes in junior high schools. She also notes that it can enhance students' interest and motivation in learning, and has advantages like enhancing language ability, being an effective learning approach, and being student centred. However, it should be noted that the activity can be repetitive, time-consuming, and put pressure on students, with some failing. Because of this, implementation should maintain the advantages and improve upon the disadvantages mentioned. She also concludes that, in the Taiwanese context, shadowing can help balance students' learning and performance by working on listening and speaking, which are often students' weaknesses.

Lin (2009) acknowledges several of the study's limitations, such as the small sample size, the short duration of the study, and, importantly, the fact that participants' responses to questionnaires and interviews may not have been full or truthful due to the impact of the researcher acting as the teacher. This is especially important to contextualise results from the interviews and questionnaires. In addition, she notes the potential of the study for the Hawthorne effect, where student participants are compelled to make a greater effort than usual as they knew their behaviour was being studied.

Overall, the study has a number of strengths. Firstly, its mixed methods design allows for effective data triangulation, with a rigorous procedure followed for the coding and categorisation of qualitative data. Secondly, the descriptions of what happened in each class, how shadowing took place, and shadowing materials are very detailed, increasing the both the trustworthiness and replicability of the study. Data collection instruments, in general, appear valid and reliable, although measurements of internal reliability and consistency for the survey, or use of validated questionnaires or items, could have improved this. There are, however, several ways in which the study could have been improved. Firstly, although the speaking test is provided as an appendix and appears to contain a range of more controlled and spontaneous tasks, there is no information provided about how the test is scored and what it is measuring in terms of speaking ability. Consultation of the test online revealed that it uses a scoring system encompassing pronunciation, intonation, and fluency, but also content relevance, and usage of grammar and vocabulary (Language Training and Testing Centre, 2016). The composite nature of the scale, and the fact that Lin (2009) does not report specific pronunciation-related scores, decreases the relevance of the study to RQ1. In addition, the issue of the teacher conducting the surveys and interviews is a significant one, and could have resulted in overly positive results had students wanted to please their teacher. Using an additional researcher to collect such data would have reduced this bias. It could also have been beneficial to involve multiple coders to reduce bias in qualitative analysis. Overall, however, this study appears relevant and appropriate to the present systematic review (especially to RQ2), and also seems to contribute trustworthy results.

19)

Martinsen, R., Montgomery, C., & Willardson, V. (2017). The effectiveness of video-based shadowing and tracking pronunciation exercises for foreign language learners. *Foreign Language Annals, 50*(4), 661-680.

Martinsen et al.'s (2017) mixed method study explored the impacts of 10 weeks of video-assisted shadowing and tracking practice on the pronunciation of French learners. Participants' attitudes to the intervention were also explored.

Participants were high school students of French, aged 15-16 and studying at a school in the state of Utah, USA. Originally, 19 students participated in the study, but one was removed as he was the only speaker of a language other than English as an L1.

During the 10-week intervention, students completed 5-10 minute tracking and shadowing activities as a class three times per week, using conversations from the textbook *D'accord!*. They also completed one 20-30 minute individual session in the language lab, shadowing videos from the *Ma France* program produced by the BBC.

Identical pre- and post-tests were used to measure pronunciation improvement. These tests consisted of one free-response picture description task, in which students spoke for 1 minute about the people and activities in a picture, and one read-aloud task, in which students read from a text containing six short dialogues.

Tests were recorded and scored by three expert raters. Raters and were blind as to whether were analysing pre- or post-test recordings. Pronunciation was scored on a scale from 1 (low intelligibility)-5 (high intelligibility), with a plus or minus when scores were slightly above or slightly below whole numbers. Raters were instructed to base scores on general accent, word stress, sentence stress, and overall intonation, and were instructed to attend to sounds like the French R and different French/English vowel sounds as precise markers at word level. They were instructed to ignore aspects of speech like fluency and grammar. Training, using a scoring guide, was conducted to increase interrater reliability, which was high in ratings of both the controlled reading (0.973) and free-response (0.971) tasks.

A paired-samples t-test was used to examine the statistical significance of differences between the pre- and post-test results. Results indicated a statistically significant improvement in students' pronunciation in the read-aloud task, with a large effect size (Cohen's d = 1.22). Student performance on the free tasks also improved, but paired-samples t tests indicated that this difference was not statistically significant.

To examine student perceptions of the intervention, weekly surveys were conducted, prompting participants about the difficulty of the weekly materials, the utility of the in-class and lab exercises, perceptions of intervention exercises. The teacher also wrote weekly journal entries to help guide analysis.

Survey responses were analysed in terms of class averages. Open responses to survey questions were categorised as positive, negative, or neutral, and subsequently analysed for emerging themes. Results showed that students were initial sceptical of the intervention (63% expressed mixed feelings and/or scepticism). However, by the end of the study 74% expressed positive views on the activities. Comments suggested that this may be because the activities got easier for the students as time went on, and that students perceived clear progress in their oral skills. Ratings of the utility of the in-class exercises showed little variation, with slight increase from average of 3 to average of 4 in second week, then average of 3.5 from weeks 3-9. Open ended responses suggested that 4 factors contributed to this: consistency, teacher feedback, audibility, and content. Students indicated it was difficult to do the activities as a group, with everyone speaking on top of each other, and felt forced to just read the subtitles. Students' perceptions of the lab activities were more positive, and improved steadily over the semester, from average of just below 4 to nearly 5. In addition, the difference in perceptions of lab activities between the first and last survey was statistically significant. Student comments suggested that the lab activities were easier to hear, more interesting, and they could go at their own pace. Overall, authenticity, technological affordances (i.e. availability of subtitles, pacing of videos, ability to manipulate speed of videos), and autonomy were identified as key themes in student comments of why they liked lab activities.

The key strengths of the study are its detailed descriptions of the procedures used, and the thorough analysis and evidencing of student perspectives on the in-class and individual practice. Use of three blind raters also adds trustworthiness to the pronunciation scores, and use of both controlled and spontaneous tasks adds ecological validity missing in many other studies.

However, there are some weaknesses that should be acknowledged. Firstly, it is difficult to untangle when students were "shadowing" and when they were "tracking", as the definitions given overlap and students seemed to have the option to do both in all activities. However, as both the activities referred to as "tracking" and the activities referred to as "shadowing" both fit under the umbrella of shadowing, as defined in this review, this does not affect the generalisability of the results for current purposes. Secondly, the study used the same task and activities for the pre- and post-tests, meaning that a repetition effect could be present in the data. In addition, whilst the authors justified the use of a free response question with reference to the literature, no other explorations of the validity of their data collection instruments were provided, for example piloting the tests and survey. Furthermore, the teacher administered the weekly surveys, and students even addressed responses to the teacher, meaning that responses may have been overly positive as students sought to please their instructor. Finally, the lack of a control group makes it difficult to be certain that pronunciation improvement was due to the shadowing and tracking activities, or just having more exposure to natural French pronunciation via the listening materials provided. In their defence, the authors are aware of this and position their study as exploratory in nature.

20)

Mishima, M., & Cheng, L. (2017). The impact of a computer-mediated shadowing activity on ESL speaking skill development: a pilot study. *L2 Journal, 9*(1), 21-35.

Mishima and Cheng (2017) conducted a pilot study to investigate the usefulness of a computer-mediated shadowing activity for improving English as a Second Language (ESL) learners' speech intelligibility. They conducted the intervention with 5 participants, 4 males and 1 female. All were Chinese graduate students at a US university who had not achieved passing scores on an in-house oral English proficiency test and were required to enrol in a course to improve speaking skills.

The activity involved the participants selecting a TED talk to then practise with imitation for two weeks, and then finally record. A novel part of the recording was that participants used the program Go Animate to create an avatar and add their sound file to a short animation with this avatar. One key issue to note about this activity, as described in the paper, is that the shadowing procedure is not described in any detail. It is simply stated that participants "repeatedly listened to the model speech to improve their mimicry" and that "they recorded and replayed their own imitated speech". There is no mention of focusing on simultaneity of speech, or on speaking whilst listening to incoming chunks.

To measure improvements in speech, two trained raters rated the final TED talk recordings based on the rubric of an in-house English test, which had a component supposedly focused on intelligibility. However, in reality this component rated comprehensibility, as it referred to listeners' ease of understanding, rather than degree of understanding. Raters also gave comments about their overall assessment of the participants' intelligibility and specific speaking problems. Scores from raters were compared to the initial scores that participants received from the in-house English test.

Comparison of scores showed a clear improvement for participants 1, 2, and 3 (from level 3 to level 4, level 1 to level 4, and level 1 to level 3, respectively). However, for participants 4 and

5, raters disagreed, with one rater giving both participants the same score as their original test (level 2), and one giving them a score one point higher for the speech sample after shadowing (level 3). Raters' comments indicated that participants 1 and 2 did not have any notable problems in their speech samples. However, comments indicated that participant 3 had some issues such as substitutions of word-final consonants, and that participant 5 was noted to be potentially unintelligible for undergraduate students.

As well as the speaking scores, an 8-item questionnaire was also conducted with participants after the intervention. Survey results from Likert scale questions suggested that students perceived the intervention as effective for improving overall speaking skills, fluency, pronunciation, rhythm, and intonation. In responses to one open question on the benefits of the activity, participants mentioned improvement in speaking skills, particularly intonation. Some participants also mentioned improvement in fluency, pronunciation, rhythm, and stress control. Survey results should be taken with caution for several reasons. Firstly, the internal consistency of the survey was not explored. Secondly, all questions are focused on improvement. Whilst participants can disagree with statements related to improvement, they were given no space to provide feedback on what they did not enjoy about the intervention, and what things it did not help them with. Because of this, the data may present an overall positive picture of students' perceptions on the intervention. Finally, the survey questions focus on the use of Go Animate, so it is difficult to entangle student views on the shadowing part of the intervention from perceptions on the use of this specific tool.

In addition, all participants participated in a 10-minute reflexive discussion (or interview) after the intervention, which primarily focused on their reactions to the use of Go Animate. Interview data was transcribed and then open-coded, which resulted in the development of over-arching thematic categories. Recurrent themes included the following. Firstly, participants commented that they found it easier to give feedback to classmates and receiving feedback themselves using avatars. They also found it more enjoyable to watch the classmate's video clips rather than just listening to recorded speech. One key issue of the reflexive discussion is that the authors do not give details on who conducted the discussions, and how they minimised pressure to give positive responses on the intervention or "please" the teacher. If this was not attended to, it could mean that the interview data presents an overly positive view of the intervention from students, as they were unable to express more ambivalent or negative views. It is also problematic that only recurrent themes in interview data were discussed, when it could have been useful to explore isolated comments from participants 4 and 5 on what they found useful or less so about the intervention.

In summary, the improvements in scores seem to provide an inconclusive picture of whether the intervention improved speech intelligibility, especially as there were only two raters and the results were not investigated for statistical significance. The usefulness of this article to a review on shadowing is also highly limited by the fact that students may not have even performed shadowing, as, although the materials could be suitable for shadowing, the procedure is not described.

21)

Mori, Y. (2011). Shadowing with oral reading: effects of combined training on the improvement of Japanese EFL learners' prosody. *Language Education & Technology, 48*, 1-22.

Mori (2011) explored the effect of a 10-week shadowing and oral reading intervention on Japanese EFL learner's prosody.

Participants were 30 Japanese second-year university students, who all received the intervention. No control group was included in the study.

The intervention included one weekly 30-minute session, in which students first listened to materials several times to familiarise themselves with content and vocabulary, then proceeded to stages of shadowing and oral reading. Materials used were five video news clips, all in American English, taken from ABC News 9. Each news clip was divided into two parts of around 1 minute and 150-200 words.

To measure improvement in prosody, a pre- post-test design was used. The pre- and post-tests were identical read aloud tests, consisting of 86 basic English words (all included in basic vocabulary for Japanese university students) combined to form phrases. Participants could practise reading before their final recording and were allowed to re-record if they made mistakes. 8 native speakers were also recorded reading the same read aloud test, to allow for comparison with a controlled model.

Acoustic analysis was used to analyse the recordings from participants and native speakers, with two clauses in particular selected for analysis due to their alternance of stressed and unstressed syllables. Waveforms, wideband spectograms, pitch contours, and intensity contours were created for the target clauses in each recording with a Kay Multi-Speech Model 3700. Durations for all syllables and segments in the two clauses were measured. Onsets and following nuclei of all stressed syllables were measured to explore any improvement in stress-related lengthening. Duration of selected syllables at end of clauses were measured to explore final lengthening. To examine speaking rates, whole duration of clauses and entire passage were measured. Mean F0 (Hz) and mean intensity (dB) values of all syllables of target clauses were measured to look at rhythmic alternation in stressed and unstressed syllables and other features of particular syllables. The pitch contour of each clause was extracted to look at improvements in intonation.

Results indicated changes in intensity, duration, and pitch from pre- to post-test. Firstly, the mean intensity of syllables was significantly greater in the post-test than the pre-test, with almost all syllables in the clauses read with significantly greater intensity at post-test. In some parts of the two clauses, increased intensity also contributed to a sharper intensity contrast between stressed and unstressed syllables. Secondly, mean durations of both the whole passage and the clauses were shorter in the post-test than the pre-test (p <0.05), indicating an increase in speech rate and/or fluency. Selected final syllables showed a significant lengthening in duration, indicating a move towards lengthening phenomena characteristic of spoken English. In addition, some unstressed syllables (such as "I" and "ing") showed a significant shortening at post-test, indicating a better control of alternative of stressed and unstressed syllables after training. Finally, pitch fell more sharply on final content words in the post-test compared to the pre-test, approaching pitch patterns of the native speakers recorded. More pairs of neighbouring stressed and unstressed syllables showed significant differences in mean F0 in the post-test (6) vs. pre-test (2), suggesting participants improved use of pitch to enhance the contrast of stressed and unstressed syllables. Overall, this data indicates that participants showed improvements in English rhythm, intonation, and final lengthening after the 10-week training.

Mori's (2011) study shows many strengths. The detailed description of the intervention is valuable for research and pedagogical application, and the acoustic analysis is an important example of how such measures can be used to track objective improvements in prosody. Some weaknesses of the study are observed by the author herself, notably the absence of a control group and difficulties disentangling the effects of oral reading and shadowing on different types of prosodic improvement. However, as the steps of the procedure described could fit with what Hamada and Suzuki (2022) described as text-presented shadowing, which is often used in combination with other forms of shadowing in pedagogical interventions (e.g. Hamada, 2021), this is not as problematic as it initially appears. However, the issue of the ecological validity of the pronunciation test is not addressed: read aloud tests may measure only controlled pronunciation, and results may not be generalisable to spontaneous speaking and real-life communication. In addition, whilst there were measurable acoustic changes in participants' pronunciation, the extent to which these changes may actually be perceivable to listeners (e.g. through adding some kind of impressionistic judgement of speech samples) was not explored.

22)

Nakanishi, Y., & Nakanishi, Y. (2015). Use of an intermediate face between a learner and a teacher in second language learning with shadowing. ***AH'15: Proceedings of the 6th Augmented Human International Conference,*** **113-116.**

Nakanishi and Nakanishi (2015) conducted a preliminary user study to explore whether shadowing with a system integrating sound shadowing and face and mouth movement shadowing (referred to as the intermediate face system) is more effective than standard shadowing in helping users improve their English pronunciation.

The intermediate face system involves a camera and a PC. The camera captures the image of both the learner and teacher, and creates an intermediate face in which the teacher's face and mouth movements are mapped onto the learner's face. The system also amplifies the learner's and teacher's mouths, so that learners can compare mouth movements in detail, and see tongue-movement. Both of these features are thought to help scaffold learners' shadowing, with the intermediate face working as a sub-goal, by showing students the mouth movements they will have to make.

The user study involved one female Japanese user. It should be noted that the user's age, English proficiency, and educational level are not reported. The user read a total of four sentences, each of approximately seven words, using normal shadowing and the intermediate face system. To control for the effect of repeating sentences on pronunciation, the methods and sentences were presented in different orders. In method sequence 1, two sentences were first read with the intermediate face system and then with normal shadowing. In method sequence 2, two sentences were first read with normal shadowing, and then with the intermediate face system. The user read each sentence for 3-5 minutes using each method, and her voice was recorded at the end of each shadowing process. The effect of video size was controlled for by ensuring that the size was the same in standard shadowing as in the intermediate face system.

12 native speaking participants were then recruited to rate the voice data, of whom three were excluded for not really being native speakers when their speech samples were taken. Participants listened to pairs of recordings for all four sentences, but were not aware of which sentences in each pair of voice data were which (i.e. standard or intermediate voice).

Participants were asked to rate which recording sounded more fluent and provide reasons for their decisions.

Rating data was analysed via frequency counts. In sequence 1, 9/18 responses indicated that pronunciation was better with the second method (standard shadowing), 6/18 responses indicated pronunciation was better with the first method (intermediate face), and 3/18 responses indicated no difference between the methods. In sequence 2, 7/18 responses indicated that the second method was preferred (intermediate face), 6/18 indicated that the first method was preferred (standard shadowing), and 5/18 indicated no difference between the methods. Reasons provided by raters judging the intermediate face recordings as superior suggested that they was better largely due to good pronunciation (12/13 responses), with good speaking speed not so commented upon (3/13 responses). Responses about the normal shadowing suggested that it was better because of good pronunciation (7/13) and good speaking speed (6/13).

From this data, the authors conclude that the intermediate face method seems to be good for improving pronunciation but could incur a high cognitive load for the user that negatively affects speaking speed. They also conclude that the intermediate face is more effective than normal shadowing. These conclusions do not entirely match the research findings. Firstly, frequency counts seem to suggest that differences between the two methods were not clear cut, with no method emerging as improving pronunciation a great deal more than the other. In addition, without statistical calculations of significance, it is difficult to make sense of which method was more effective, and whether or not this was due to chance. It does seem that raters perceived the intermediate face recordings as having better pronunciation but worse speaking speed, compared to the recordings from standard shadowing. However, this finding is also difficult to interpret without statistical analysis. Overall, these issues, combined with the fact that there was only one user in the study, make the trustworthiness of the findings low. In their defence, the authors do acknowledge that the study is not a rigorous or quantitative experiment. However, in this case they should have been more tentative in the interpretation of their findings.

23)

Nakayama, T. (2021). Effectiveness of the visual-auditory shadowing method on learning the pronunciation of kanji. *Japanese Psychological Research, 63*(1), 26-34.

Nakayama (2021) investigated whether Visual Auditory (VA) and Visual Visual (VV) shadowing can help students of Japanese learn the pronunciation of kanji.

She conducted a study with 95 immersion students of Japanese from the first to seventh grades (ages 6-12). The students were from the USA, and spoke English as an L1.

Participants were divided into VA intervention groups (n = 48) and VV intervention groups (n = 47), and these groups were then divided into three literacy groups (level 1, level 2, level 3). The VA or VV interventions both trained participants to learn 12 kanji sets, with the VA intervention presenting the kanji and an audio recording of its pronunciation, and the VV intervention presenting the kanji and its hiragana pronunciation.

To measure pre-existing knowledge and improvement of kanji pronunciation, a 14-item test was used, in which the 12 target kanji were tested, along with two dummy items. Participants

received the pre-test, either the VA or VV intervention, to learn the pronunciation of kanji, and a post-test in the same session. They also received a delayed post-test one week later.

The pre- and post-test scores were compared, and ANCOVA were run to reduce the impact of preliminary knowledge of kanji. ANCOVA were run to investigate effect of independent variables (literacy level and training method) on post-test, then delayed post-test. Results showed that the VA method was significantly more effective than the VV method as a whole, although overall improvement was small. In addition, no difference in efficacy was found in the VA and VV conditions in Level 1 and Level 2 students. Nakayama (2021) concludes that low level learners may have too low a proficiency in reading and listening skills to benefit from either intervention, but that the VA intervention may better facilitate kanji learning amongst those students who have reached a higher level (Level 3 in the study).

Whilst the experimental design and results seem robust in assessing whether learners are able to learn the pronunciation of new kanji, they are limited by the lack of control group. In addition, they do not seem relevant to this review due to the way in which shadowing is used. In the VA training, learners are asked to complete what are essentially listen and repeat activities of single words. This removes the element of shadowing based on speaking one part of a phrase whilst listening to the next. In the VV training, there is no element of shadowing as it is traditionally defined, as there is no repetition of auditory input. For these reasons, the study was not included in the results section of this dissertation.

24)

Nakayama, T., & Armstrong, T. (2011, November). ***Weak forms in shadowing: how can Japanese EFL learners perform better on shadowing tasks?*** **[Paper presentation]. Japan Association for Language Teaching (JALT) International Conference, Tokyo.**

Nakayama and Armstrong (2011) explored the effect of two different priming methods, visual-auditory shadowing (VA) or scaffolded auditory shadowing (SA), on participants' learning of weak forms of function words.

95 first year Japanese L1 students from a Japanese university participated in the study. They were divided into two groups, one of which received VA intervention, and the other of which received the SA intervention. The VA intervention involved matching visual input with auditory input at natural speed. However, this is the only information given about the intervention, which should have been described in more detail. The SA intervention involved three priming steps: first using very slow, enunciated auditory input, then moving on to careful pronunciation, and ending with relaxed pronunciation as the final step. Assessment before the intervention revealed no significant differences in listening ability between the two groups. The interventions were conducted over approximately 6 weeks, between the beginning of June and mid July (2011). The paper does not specify how many training sessions participants received in total, but notes that each session was 60 minutes long, and that all sessions were conducted by the first author.

To measure participants' improvement in shadowing weak forms, they were asked to shadow the same speech as a pre- and post- test. Two of the authors then listened to the recordings and counted the number of content and function words successfully shadowed, before running a MANOVA test to analyse the significance of the main effects of groups and tests.

The MANOVA results showed that the overall main effect for both groups was not significant ($p > .05$), and that the main effect for the groups was significant for function words only ($p < .05$). In addition, the VA group outperformed the SA group in the post-test ($p < .01$), and the simple main effect for the tests was significant for the VA groups. Finally, the increased amount of function words was significantly larger than that of content words in the VA group.

The authors conclude that VA facilitated the production of weak forms in greater quantity than scaffolded prime. Therefore, it can be an effective way to demonstrate differences between clear speech and weak form of function words. They hypothesise that VA potentially worked better than scaffolded auditory shadowing for three reasons. Firstly, VA participants had repeated exposure to the same prime, whereas for the SA group auditory quality and speed changed each time. Because of this, SA participants may not have been able to match words to input as effectively. Secondly, in the SA group, learners' phonological knowledge may not have matched even the enunciated auditory input, which could have created confusion and processing difficulties for students. VA participants, on the other hand, had 3 chances to adjust their phonological knowledge for natural speed as they received same input at natural speech three times. And finally, as the speed was stable in the VA group, VA participants could more easily control their cognitive attentional resources to focus.

Several key limitations of this study are not addressed in the article. Firstly, the same speech was used for both the pre- and post- tests and the shadowing training. This could have introduced a repetition effect into the results, and also does not allow for any exploration of whether learners were able to generalise what they learnt. Secondly, there was no control group, which would have allowed the researchers to explore the real effect of the interventions in a more robust way. Finally, use of a shadowing task as a pre- and post-test may not provide the most ecologically valid picture of improvement in pronunciation of function words: a spontaneous task could had added to validity in this regard.

25)

Nguyen, H., & Nguyen, M. (2019). Applying shadowing technique and authentic materials to promote phonological awareness amongst young learners of English. *Proceedings of ELT Upgrades 2019: a focus on methodology*, 13-23.

Nguyen and Nguyen (2019) conducted a mixed-method study to explore the impacts of combining shadowing training and authentic materials on young learners' phonological awareness of English.

Participants were 40 young learners in grade 5, who had a proficiency of between A1-A2 and were L1 speakers of Vietnamese. The country in which the study was carried out is not stated, but from participants' L1 is assumed to be Vietnam. Participants were studying English at a language centre as an extracurricular activity, and received around 4 hours instruction per week.

All 40 students participated in the shadowing intervention, which lasted 8 weeks. During this time, participants shadowed a total of 20 audios and videos, sourced from YouTube and dealing with topics related to daily communication. The length of the materials used, and the exact procedure that was followed to work with them, is not stated.

Data collection instruments used were a diagnostic assessment (testing participants' fluency and coherence, grammatical range and accuracy, lexical resource, and pronunciation), a portfolio in which the teacher recorded students' strengths and weaknesses, and a survey to explore students' attitudes to the intervention.

However, data from these instruments does not appear to be shared, as the paper seems to rely on data from a pre- post-test of phonological awareness and/or pronunciation performance that is not specified. As such, its reliability and validity cannot be gauged.

The data from the pre- and post-test is presented with calculations of raw percentages of students that improved in specific features of pronunciation: exponents, naturalness, idea clarity, intonation, stress, connected speech, phonetic variants, silent sounds, consonants, and vowels. It is unclear how performance in these dimensions was measured, and how and by whom scoring was calculated.

Results indicated that 65% of participants showed overall improvement at post-test, with 15% showing no change, and 20% decreasing their score. A paired-samples t-test indicated that student improvement was significant. In addition, phonological awareness is reported to have improved on more specific measures, with 85% of students improving in stress, 80% improving in intonation, 72.5% improving in exponents and consonants, respectively, 70% improving in vowels, 67.5% improving in phonetic variants, 60% improving in connected speech and silent sounds, and 55% improving in naturalness. It should be noted that no tests for statistical significance of these individual results were performed.

From these results, the authors conclude that the study "proved" the effectiveness of the combination of shadowing and authentic material on young learners' phonological awareness. This is a completely unfounded claim, considering the significant issues present in the study related to lack of reporting. With no information given on data collection instruments and analyses (i.e., the tests used, how they were validated and tested for reliability, the scoring system, interrater reliability) the results cannot be considered trustworthy, let alone "proof" of the effectiveness of the intervention. Failure to describe key aspects of the population, with the country in which the study was carried out not mentioned, and participants' L1 only mentioned in the discussion, compound this problem. In addition, the lack of statistical tests when exploring non-composite measures is problematic. Even if these aspects had been addressed in a satisfactory way, the lack of control group would have limited the generalisability of the results.

As such, whilst the research questions and design are highly relevant to this systematic review, this study should be taken with caution.

26)

Omar, H., & Umehara, M. (2010). Using a "shadowing technique" to improve English pronunciation of deficient adult Japanese learners: an action research on expatriate Japanese adult learners. ***The Journal of Asia TEFL, 7*****(2), 199-230.**

Omar and Umehara (2010) conducted a 6-month action research project (from September 2006-February 2007) to investigate whether shadowing would be effective at improving Japanese participants' rhythm when speaking English. The study was conducted with four retired expatriate Japanese adult learners who were living in Malaysia. The participants were

selected via purposive sampling, as they had problems with English pronunciation and had been taking English classes out of necessity since arriving in Malaysia. Data was collected via recorded tapes of participants' shadowing, learner journals, a questionnaire about participants' perceptions of their ability and participant observation notes.

The study involved five action research cycles. Before starting the cycles, researchers asked participants about the situations in which they had trouble communicating with people in Malaysia in daily life. Participants indicated a range of daily situations and that they were often not understood due to their pronunciation. The situations were then used to design the teaching instruction in the main project.

In the first cycle of action research, a pronunciation test was administered to determine the weakness of each participant. Key difficulties included rhythm, TH sounds and /l/ and /r/ sounds. Next, shadowing was implemented once a week for 20-30 minutes. Learners' pronunciation whilst shadowing was recorded every two lessons. Learners were initially discouraged by the shadowing practice as they found the content too difficult to understand and imitate, which lead to the researchers substituting the material for the second cycle.

The second cycle involved shadowing practice with easier materials and slow rates of speech, as well as lessons and controlled practice of the difficult sounds identified in the pronunciation test. In this second cycle, researchers observed that participants struggled to produce "simplifications" of words like weak forms or connected speech.

As such, the third cycle of action research involved teaching simplifications in detail and continued shadowing practice. Researchers observed that participants were beginning to note they could use some of what they were shadowing in daily conversation.

In cycle four, participants shadowed more challenging materials with a faster rate of speech, which was difficult for them. Because of this, researchers allowed participants a short pause before imitating each part of the recording.

Participants also found it challenging to "catch" chunks of speech, and to stress syllables, which lead to instruction on stress, rhythm and intonation in cycle five. Cycle five also challenged learners to shadowing without any pause to repeat.

Overall, learners reported progress in shadowing speaking and listening in their diaries. They also reported that they had difficulty shadowing unfamiliar vocabulary and the flow of connected speech. In their diaries, some reported mixed feelings about shadowing at the beginning (fear, scepticism, confusion), but all seemed more enthusiastic about and interested in it by the end. From observation and analysis of shadowing recordings, the researchers say that all participants' sense of rhythm improved greatly throughout the intervention. Analysis of the shadowing recordings also showed that participants were able to reproduce more speech from a 2 second frame when shadowing at the end of the cycle. Participants also reported, via a questionnaire conducted at the end of cycle 5, that they could produce good English rhythms and read or pronounce English faster.

However, these results should be taken with caution as no data or excerpts from the diaries, analysis of recordings, or participant observation is presented explicitly or as an annex. In addition, self-reporting of improvement of rhythm and fluency in the questionnaire could be problematic.

27)

Ono, Y., Ishihara, M., & Yamashiro, M. (2012). Mobile-based shadowing materials in foreign language teaching. *The 1st Global Conference on Consumer Electronics 2021*, 90-93.

Ono et al. (2012) conducted two studies to explore the effect of a mobile-based shadowing course on students' oral reading ability and attitudes to English language learning.

In the first study, participants were 34 second year students at a technology college in Japan, assumed to be 16 years old from online consultation of how the school system works in Japan. It should be noted that no information is given about participants' English proficiency, gender, or L1. Participants took a 4-week course in which they shadowed materials from a film on iPod Touches.

A test of oral reading ability was used as a pre- and post-test. The test was created from third grade Society for Testing English Proficiency listening questions and included monologue passages with an average of 35/7 words and a 5.1 Fresch-Kincaid Grade Level. Whilst the grade level does suggest that the test was of an appropriate level for students, it should be noted that there is no other mention of how the authors addressed the validity and reliability of the test, for example, through piloting. There is also no mention of whether the same passages were used as pre- and post-test. This is important as, if they were, there may have been a possible test, re-test effect and, if they were not, the authors would have needed to take measures to ensure that the two tests were of a comparable difficulty level.

Oral reading scores were calculated via 16 assessment categories (word level: pronunciation, stress, intonation, attitude, volume; sentence level: pronunciation, stress, sentence stress, intonation, fluency, liaison, sense reading, attitude, volume; holistic evaluation: total impression). Raters scored these categories on a 5-point Likert scale. It is, however, unclear how many raters took part in analysis, and no calculations of interrater reliability are given. In addition, it is unclear whether raters were blind with respect to whether they were rating pre- or post-test recordings.

After rating, t-tests were used to explore differences between pre- and post-test results. Factor analysis, with unweighted least squares method with Varimax rotation, was then carried out. Results indicated that all criteria for oral reading showed significant improvement between pre- and post-test, and that 3 main factors contributed to the development of oral reading skills: "natural flow of utterance" (related to categories "stress" and "fluency"), "volume" and "intonation".

In the second study, 76 fourth year Japanese high school students, studying at the same college and assumed to be age 18, undertook a 7-week shadowing course of the same nature as that of the first study. Once again, no information is given about participants' L1 or English proficiency. A questionnaire about attitudes to English learning was conducted as a pre- and post-test. It included 18 items to which participants responded on a 5-point Likert scale.

T-tests were used to explore differences between pre- and post-test scores. Factor analysis, with unweighted least squares method with Varimax rotation, was then used. Results indicated that

nine questionnaire items showed significant improvement between pre- and post-test, with the main factors of changed consciousness being "attitude towards iPod Touch", "learning strategy" and "attitude towards active communication".

The authors conclude from their results that the iPod Touch shadowing course was effective in both improving oral reading skills and changing attitudes towards language learning. They note that mobile-based programs may offer new possibilities inside and outside the classroom.

This study's strength lies in its detailed description of the oral reading scores rubric, which clearly shows the many dimensions of pronunciation measured in the study word level: pronunciation, stress, intonation, volume; sentence level: pronunciation, stress, sentence stress, intonation, fluency, liaison, volume). This makes it relevant to the present systematic review.

There are, however, a number of weaknesses of the study that limit the generalisability and trustworthiness of its findings. First, the populations of the study are not described in sufficiency detail. L1 and proficiency level could have significant impacts on the benefits from shadowing, and are therefore important to state. In a similar vein, the procedure involved in rating the oral reading recordings is also not described in detail, with important issues like numbers of raters, calculations of interrater reliability, whether raters were blind with respect to pre- and post-test recordings, and whether pre- and post-tests were the same or different, glossed over. Data collection instruments should also have been validated and tested for reliability in a more rigorous way, for example via piloting or use of measures validated in previous studies. A more ecologically valid test of pronunciation, including a spontaneous measure, would have also been beneficial. Finally, the study design could have been improved with the addition of a control group, and with the addition of a qualitative component, such as semi-structured interviews or classroom observations, to explore student perceptions in greater detail.

28)

Rojczyk, A. (2013). Phonetic imitation of L2 vowels in a rapid shadowing task. In J. Levis & K. LeVelle (Eds.), ***Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference.*** **Aug. 2012. (pp. 66-76). Ames, IA: Iowa State University.**

Rojczyk (2013) investigated if and to what degree a shadowing task would allow Polish learners to approximate target-like formant frequencies of the English vowel /æ/ . The study design was quantitative and cross-sectional, and involving one 20-minute lab training session.

Participants were 22 native speakers of Polish from a Polish university (16 females, 6 males). Their average age was 19.8 and their English proficiency ranged from intermediate to upper intermediate.

In the lab session, participants first read a word list including 12 monosyllabic target words, which included the vowel /æ/ flanked by consonants, and 12 foil words. Their productions were recorded to establish a baseline of pronunciation. Next, participants imitated the same words after a model speaker (a male southern British English speaker). Participants were instructed to repeat as immediately as possible, and the presentation of words was separated by a two-second interval. Again, productions were recorded to allow for comparison in

analysis. Finally, participants read /bVt/ sequences with Polish vowels /i, e, a, o, u/ that were used to help establish the acoustic space for each talker in normalisation.

From the recordings, formant frequencies of vowels were measured at vowel mid-point using Praat. In order to compare the distance between participants and the model speaker, anatomical and physiological variation between the participants was normalised using the Lobanov transform. Then Euclidean distance was computed between participants' and the model's F1 and F2 frequencies. Finally, data was analysed via a two-way mixed ANOVA with task (word list, imitation) and gender (male, female) as independent variables, and Euclidean distance as dependent variable. Scatter plots were used to inspect the clustering of participants' vowels with the model vowels.

In terms of clustering, shadowed productions were more centred around the model speaker's productions. Unlike vowels from the word list, shadowed vowels were also characterised by less extreme productions towards either Polish /e/ or /a/. So, the model auditory input seemed to create a magnet effect by cancelling less extremely outlying productions. Analysis of Euclidean distances showed a highly significant main effect for task on the magnitude of convergence, with participants modifying productions of the target vowels to approximate the model in imitation compared to baseline word reading. There was no significant gender task interaction.

From this data, the author concludes that foreign language learners are able to modify their productions of non-native vowels as a result of exposure to a model.

The study design and analysis are described in detail, and are appropriate for their research questions. However, the research presented here does not fit with this systematic review as the task used is not representative of how shadowing is defined in the literature. It is rapid imitation of words, not rapid repetition of sentences combining simultaneous speaking and listening to incoming information. It should be noted that the author does not make any reference to literature when defining shadowing, simply defining it as "characterised by a minimum time-lag between hearing the model and actual imitation" (p. 67). For these reasons, this study is not discussed in the results section of this dissertation.

29)

Rongna, A. & Hayashi, R. (2012). Accuracy of Japanese pitch accent rises during and after shadowing training. *Proc. Speech Prosody 2012*, 214-217.

Rongna and Hayashi (2012) conducted a longitudinal study to investigate the effect of a 7-week shadowing training program on Japanese learners' speech rate and accuracy of pitch accents. They also explored the interaction of the training with proficiency level.

Participants were 15 students of Japanese as a foreign language, recruited from a language school in Kobe. 11 were native speakers of Chinese, and four were native speakers of Mongolian, and all had been studying Japanese for between 0.5 and 3.5 years. Participants' proficiency, overall, was of an intermediate level (around Japanese Language Proficiency Test level N2).

Participants were divided into a higher (n = 9) and lower (n = 6) proficiency group, based on their scores on a test administered by the researchers. Both groups received 7 weeks of shadowing training. The training consisted of three sessions in week one, two, and seven, in which participants shadowed a dialogue 10 times per session without seeing the script. Definitions of shadowing appear in line with those of this literature review, i.e. shadowing sentences simultaneously as possible and therefore speaking whilst listening to incoming information. It appears that participants shadowed the same 227 morae dialogue in all training sessions. The dialogue was chosen through discussion with teachers of the language school to ensure it would be suitable for both the lower and higher proficiency groups. However, it should be noted that no other information is given about validation of the test, e.g. via piloting.

Participants read the text they shadowed at three time points during the study: before the intervention, in week two (after shadowing) and in week seven (before shadowing). Participants' reading was recorded, and these recordings were analysed for improvement in speech rate and pitch accuracy.

Acoustic analysis using a speech analyser (Wavesurfer) was used to measure speech rate. The pattern of pitch accent of 23 nouns was determined by the second author, who evaluated whether participants' productions of pitch accent were identical to the model recording they heard. It should be noted that, as only one rater was involved, and as no information is given about whether rating was blind with regard to time point, results related to pitch accent should be taken with some caution. Two-factor repeated measure ANOVAs were used to explore the significance of improvements in speech rate, pitch accent, and patterns of pitch accent.

Results for speech rate identified a significant main effect for task (1-4), indicating groups made improvement in speech rate over time. There was also a pronounced improvement from time point one to time point two. However, there was no significant main effect for group, meaning the proficiency was not related to improvements in speech rate.

Analysis of improvement in pitch accent identified a significant main effect for task, indicating that groups made improvement in pitch accent over time. However, there was no significant difference between the two groups, which once again indicated that proficiency was not correlated with improvement. Pitch accent accuracy improved dramatically at the second reading and remained high for both groups from this point on. Bonferroni-corrected post hoc comparisons indicated that there was a significant difference between reading 1 and 2, and reading 1 and 4.

In terms of patterns of pitch accent, words were of type 0 accent (flat accent without nuclear) and 4 words were type 1 (accent nuclear on first mora). For type 0 accents, analysis indicated that there was a significant main effect for task, but not for language. For type 1 accents, however, Chinese speakers showed a tendency to have higher accuracy than Mongolian speakers, with Chinese speakers showing 100% accuracy from the second reading onwards. All four groups (Chinese, Mongolian, lower, higher proficiency) demonstrated significantly higher accuracy of pitch accent at reading 2 and 4 than reading 1.

From these results, the authors conclude that longitudinal shadowing training helps learners of Japanese, both those of lower and higher proficiency, to successfully realise pitch accents and improve speech rate. They note that Chinese speakers improved more with pitch accent accuracy than Mongolians, and that improvement in patterns of pitch accent was related to proficiency.

The conclusion that the shadowing training improved speech rate seems well founded, as it is based on acoustic analysis from a computer program. However, the conclusions related to pitch accent should be taken with more caution, as only one rater, one of the researchers themselves conducted the analysis, and no information was given about whether the rater was blind as to whether recordings were from time 1, 2, or 3. For this reason, there may be an element of subjectivity and bias in the data. In addition, the authors conclude that proficiency played a role in accuracy of pitch accents, but cite no data in text to support this, with data only appearing on a graph. They also do not have seemed to have performed any ANOVA analyses on this data to check for its statistical significance.

There are also a number of other features of the study that limit the generalisability of its findings. Firstly, the lack of control group. Secondly, the sample size was very small and unbalanced, making conclusions about the effects of proficiency and L1 very tentative. It seems unfounded to draw firm conclusions on Mongolian learners' difficulties with pitch accent, for instance, on the basis of the results of four participants. The authors also do not provide information on participants' age, educational level, gender, or length of residency in Japan, which all could have acted as confounding variables. Third, it should be noted that, as the only data collection instrument used to record pronunciation improvement was a read aloud dialogue (the same one the students shadowed in every practice session), results may not be generalisable to overall pronunciation improvements, for instance in spontaneous speaking. The ecological validity of this study is therefore highly limited. Another speaking task should have been used to deal with this limitation.

30)

Rongna, A., Hayashi, R., & Kitamura, T. (2013). Naturalness on Japanese pronunciation before and after shadowing training and prosody modified stimuli. *SLaTE 2013*, 143-146.

Rongna et al. (2013) conducted a study to investigate naturalness judgements of participants' speech after shadowing training, and the factors that may have contributed to improvements in naturalness. Data was used from a study that the authors had previously conducted.

Participants were 33 Chinese and Mongolian learners of Japanese as Foreign Language, who were majoring in Japanese at a university in Inner Mongolia. Their proficiency was of intermediate level, and all had learnt Japanese for three years. Participants were divided into two groups, one of which received shadowing training ($n = 19$) and the other of which received repeating training ($n = 14$). Texts were the same short phrases for both groups, read by a native male Japanese speaker. The texts were the same at pre- and post-test, and during training. Participants' utterances were recorded at pre- and post-test, and utterances from four learners from each group (two Chinese, two Mongolian) were selected for analysis in this study, on the basis that their naturalness judgements were almost identical at pre-test.

The pre- and post-recordings from the eight selected learners, as well as three types of modified stimuli (modified for duration, pitch, and duration and pitch), the model speech, and five dummies were rated for naturalness. There were 46 stimuli in total, which were randomised and evaluated by 52 native Japanese speakers (26 Kansai dialect speakers, 26 Tokyo dialect speakers). Raters used a Likert scale in which 1 was extremely unnatural and 7 was extremely natural or native-like. All raters practised with five stimuli before completing the task. Statistical analyses were used to explore difference in ratings of the different stimuli.

Results indicated that there were no significant differences between Kansai dialect speakers and Tokyo dialect speakers for the five types of stimuli, so all scores from all raters were included for analyses. Two-way analysis of variance with group (shadowing and repeating) and type of stimuli (pre, post, dur, F0, durF0) showed a significant main effect in type of stimuli (p<.001) and marginal significance in interaction (p = .06). The results of multiple comparisons of Bonferroni test indicated that the scores of durF0 were significantly higher than others in both training groups (p<.05). For stimuli at post-test, the score of the shadowing group was significantly higher than that of the repeating group (p<.05). Analysis of durational pattern of the natural pre- and post- stimuli indicated that, at post-test, the mora deviation from model speech became significantly smaller (p>.05). Analysis of word accent pattern indicated that the shadowing group tended to show higher accuracy than the repeating group at post-test, although this difference was not significant. The mean score of dur was higher than that of the post in the repeating group (p<.05).

From these results, the authors highlight that the shadowing group received higher naturalness judgements than the repeating group, potentially because of improvements in the moraic duration and the accuracy of pitch accent. In addition, as neither durational pattern nor F0 pattern alone caused significant differences in the naturalness judgements of modified stimuli, they conclude that it is important to deal with both features of speech together in JFL pronunciation instruction.

The study, overall, appears to be robust. However, the generalisability of its results are limited by the small number of recordings rated and analysed (from only eight learners) and the tasks used to collect data on pronunciation. As the oral text was the same at pre-test, post-test, and during practice, a large degree of improvement in both groups could be due to a repeat effect. In addition, the ecological validity of such a highly controlled task is limited. To provide more generalisable measurements of pronunciation improvement, it would have been beneficial to include a task that 1) differed from the pre-test and training, and 2) prompted some form of spontaneous speaking from participants. In addition to these points of improvement, it should be noted that the duration of the shadowing and repeating training is not provided, limiting comparison with other studies.

31)

Rongna, A., Hayashi, R., & Kitamura, T. (2015). Crucial prosodic features in Japanese learners' pronunciation: evidence from naturalness judgments of synthetic speech. ***Journal of the Phonetic Society of Japan, 19*****(3), 37-42.**

Rongna et al. (2015) investigated the impact of shadowing and repetition training on naturalness of participants' pronunciation. They also investigated the features of speech which contributed to judgements of naturalness.

Eight participants took part in the study, all of whom were intermediate learners of Japanese as a Foreign Language who had never been to Japan. Two males and six females participated in the study. The age, educational level, or L1 of participants is not provided.

Four participants undertook shadowing training, in which they were instructed to imitate sentences as simultaneously as possible, without seeing the script. Four participants undertook

repetition training, in which they were instructed to repeat sections of sentences without seeing the script. Both groups used the same speech models. It should be noted that no information is provided about the duration of the training for either group, whether both groups were taught by the same teacher, and whether materials used for training were the same as those used for testing.

Pre- and post-tests were conducted to explore improvements in naturalness. These tests were identical, and involved reading of 15 model sentences. Recordings of participants' speech were taken, and a model sample of a native speaker of Japanese reading the same sentences was used for comparison.

Before judgements of naturalness and analysis of prosodic features contributing to naturalness, Praat was used to synthesize the pre-test recordings into 3 types of stimuli: duration (DUR), pitch (PIT), and duration + pitch (DUR + PIT). Other modifications were made to allow comparability with the model speech, for example time-warping of phonemes.

After this, pre and post utterances, as well as the three types of synthesized stimuli and the model speech were used for naturalness judgments. In total, 46 stimuli were randomised and evaluated by 70 native Japanese speakers. 44 raters were Tokyo dialect speakers, and 26 were Kinki dialiect speakers. The raters had to assess the naturalness of stimuli using a 7- point Likert scale (1 = very bad, 7 = very good). 5 dummy stimuli were used for practice before the assessment.

In addition, pre- and post-test recordings were analysed for duration of each mora, with the deviation of the moraic pattern from the model speech calculated for each participant. To investigate pitch accent, two phonetically trained Japanese native speakers assessed whether the pitch patterns of participants were the same as those of the model speech. They wrote down the accent nucleus of 13 words in the recordings, then the accuracy of pitch accent was calculated (number of accurately pitched words/13). A two-way repeated measures ANOVA was performed between the groups and the stimuli (pre, post, duration, pitch, duration + pitch) to investigate what prosodic feature was important for naturalness.

Results indicated that there were no significant main effects for evaluator's dialects, so mean scores of all 70 judges were used for further analysis. Overall, naturalness scores of the shadowing group were significantly higher than those of the repeating group at post-test. In terms of mora duration, the standard deviation of mora duration in the participants' speech before training was 50ms in the shadowing group and 45.5ms in the repeating group. After training, the standard deviation was significantly reduced in the shadowing group, but not the repeating group. The deviation of the shadowing group was significantly smaller than that of the repeating group at post-test. The accuracy of word accent showed no significant difference between shadowing and repeating groups or between pre- and post-test. However, the shadowing group did tend to show higher accuracy than the repeating group by about 10 points. Judges assessed the stimulus DUR + PIT, whose phoneme duration and pitch frequency patterns were modified to be the same as the model speech, as most natural.

The authors conclude that shadowing was more effective than repeating in improving naturalness scores and in improving mora timed rhythm. They hypothesise that this could be because shadowing is an online process, which could potentially effect the durational structure of participants' pronunciation more easily than the offline processes involved in repetition. They also hypothesise that DUR may have contributed more to assessments of naturalness as

pitch in pre-test recordings was already quite accurate, meaning that synthetic improvements of pitch may not have increased feelings of naturalness.

Overall, this is an interesting study that begins to explore shadowing in the improvement of Japanese pronunciation. It is particularly rigorous in the number of raters used and in its isolating and analysis of the exact acoustic features contributing to naturalness. However, the study should be treated as exploratory due to the small sample size, lack of control group, and lack of detailed descriptions of both participants (most crucially, their L1 and age) and the training methods used (most crucially the duration of the intervention, and whether the same teacher was used for both interventions). In addition, the only measure of pronunciation is a controlled practice exercise, which should have been supplemented with a spontaneous speaking exercise to provide a fuller picture of pronunciation improvement. All of these factors severely limit the generalisability of the findings.

32)

Saito, Y., Nagasawa, Y., & Ishikawa, S. (2011). Effective instruction of shadowing using a movie. In A. Stewart (Ed.), *JALT2010 Conference Proceedings* (pp. 139-148). JALT.

Saito et al. (2011) conducted a shadowing intervention with 41 third year high school students of English in a high school in Japan. The intervention took place over three classes and used two conversations from *Charlie and the Chocolate Factory* as shadowing material. Shadowing involved a 6-step process based on Kadota and Tamai (2004), including dictation, synchronised reading and understanding meaning, prosodic shadowing, checking weak points and content shadowing. After the intervention, Saito (2011) administered a questionnaire in which students were asked about their perceptions of shadowing.

Survey results showed a range of findings. 85% of students knew what shadowing was. 93% had done shadowing before. When asked what points they were aware of whilst shadowing, 61% stated rhythm, 44% stated pronunciation, 11% stated accent, 12% stated intonation, and 10% stated stress. 41% of students believed they had done well in the shadowing intervention, and 59% stated they had had trouble. 38% thought the intervention was fun, whilst 48% found it difficult. When responding to what kind of skill is developed by shadowing, 90% replied listening and speaking, and 36% stated reading. When asked how often they would like to practise shadowing, 88% replied at least once a week. When asked about what materials they would like for shadowing, 70% specified films. Analysis of open responses to questions showed that some students believed shadowing was a good exercise to practise listening skills and improve pronunciation and oral English.

Saito et al. (2011) conclude that shadowing may help Japanese students raise their awareness of the rhythm of the English language in an enjoyable way. The first of these claims seems reasonable, as students did indeed state that they became more aware of rhythm when shadowing. However, as only 38% enjoyed the activity, the second claim may be overstretched. It does seem that students may like to shadow with films. It is worth noting that, as none of the survey questions appeared to contain an option to select “none of the above” or “other” responses should be taken with caution. In addition, it is unclear who conducted the survey with students: if it was their teacher positive results may be somewhat inflated as students may have wanted to please their instructor, or have been wary about negative consequences of expressing critical views.

33)

Shao, Y., Saito, K., & Tierney, A. (2023). How does having a good ear promote instructed second language pronunciation development? Roles of domain-general auditory processing in choral repetition training. *TESOL Quarterly, 57*(1), 33-63.

Shao et al. (2023) investigated the effect of two weeks of shadowing training on students' comprehensibility and accentedness. They also explored the extent to which participants' perceptual acuity and audio-motor integration were correlated to improvements.

Participants were 47 Chinese high school students in Chengdu, aged 17-18. The study included 22 males and 25 females, all of whom had been studying English for around 6 years, and none of whom had received any prior pronunciation training. The participants were divided into an experimental group (n = 37) and a control group (n = 10).

The experimental group received two weeks of training, including 12 30-minute shadowing sessions. Shadowing materials were selected from the application English Fun Dubbing, which presents short segments of videos with line-by-line transcripts. Instructions were given in Chinese, and participants then completed their shadowing practice with headphones and iPads. Participants were separated in the classroom and the corridor to minimise the impact of background noise. The control group, on the other hand, received grammar, vocabulary, and writing training instead of the shadowing practice.

All participants completed two pre- and post-tests, in week 1 and week 4, respectively, to measure improvements in accentedness and comprehensibility. One test measured controlled speaking, and involved reading six sentences selected from the shadowing materials. This test was identical at pre- and post-test. The other measured spontaneous speaking, and was a picture description task. For this task, two different versions of the test were counter-balanced across group and time, to minimise the impact of topic on results. All pre- and post-tests were recorded, resulting in 188 recordings (47 participants*2 tests*2 testing points).

Recordings were then analysed by five raters, all of whom were advanced L2 speakers of English and graduate students in applied linguistics. Raters were given detailed explanations of the constructs of accentedness and comprehensibility, and rating guidelines. To reduce rater fatigue, only the first 30 seconds of the speech samples were used. Speech samples were rated on a 9-point scale for comprehensibility and accentedness. Inter-rater reliability was quite high (.88 for comprehensibility and .85 for accentedness).

Independent t-tests showed that the experimental and control groups' performance was comparable for most contexts at pre-test (comprehensibility in the spontaneous task, accentedness in the controlled task, accentedness in the spontaneous task). However, the experimental group had a higher comprehensibility score than the control group and the group difference was marginally significant. A multiple comparison analysis showed that the experimental group significantly improved their comprehensibility scores over time on both the controlled task and the spontaneous task, with a medium to large effect size. The control group's gains in comprehensibility did not reach statistical significance on either task. For accentedness, a multiple comparison analysis showed a medium effect of improvement for the experimental group on both controlled and spontaneous tasks. The control group's performance on accentedness did not change significantly over time.

Tests for auditory processing ability were administered in weeks 2 and 3. The auditory-motor integration test included tests of rhythm reproduction (10 rhythmic patterns which participants had to repeat) and melody reproduction (10 melodies which participants had to repeat). The auditory acuity test consisted of three subtests designed to measure the ability to perceive the spectral and temporal details of a sound, including formant, pitch, and duration discrimination thresholds. Composite scores were calculated for auditory acuity (using standardising and averaging for formant, pitch and duration discrimination) and integration (standardising and averaging rhythm and melody reproduction). Partial correlation analyses, controlling for pre-test scores, showed that gain scores for accentedness were significantly related to individual differences in audio-motor integration. However, the link between acuity and accentedness did not reach statistical significance, and no significant correlations were found for comprehensibilty.

From these results, the authors conclude that shadowing can be an effective method of improving pronunciation in a short period of time, particularly in terms of comprehensibility but also in terms of accentedness. They also highlight that their findings provide support for the potential of mobile-based assisted language learning in classrooms where language exposure is limited in quantity and quality. However, they suggest that students with high audio-motor integration ability may benefit more from shadowing, as it echoes the same skills tested in rhythm and melody tests. They suggest that other students may benefit more from other types of training, and that this should be explored in more detail.

Shao et al.'s (2023) work is a rigorous piece of research that has several notable strengths. Firstly, they provide a detailed description of the procedures and materials involved in shadowing and also those followed by the control group. They also control for the influence of a number of variables in their study, for example prior pronunciation training, use of shadowing outside of class, and the impact of test and re-test through counterbalancing picture tasks. Secondly, they use both controlled and spontaneous tasks to gather speech samples, giving the study a level of ecological validity not found in many others in the sample for this review. Their use of five raters with prior training is also speaks to the reliability of analysis.

There are some potential issues with the study that are unclear in the author's description of their methodology and analysis. Firstly, it is not mentioned whether the control and experimental groups were taught by the same teacher. As the teacher is a particularly powerful variable in such quasi-experimental studies, this information should have been provided. Secondly, it is unclear whether rating was blind with respect to pre- and post-tests. The authors do mention that controlled speaking samples were analysed before spontaneous samples, but do not provide information with regard to how they reduced rater bias with regard to pre- and post-test recordings. Finally, the study has a number of limitations which the authors themselves acknowledge, namely the fact that the use of sentences from shadowing materials may have given an advantage to the experimental group, and the fact that the number of students in the control and experimental group was highly unbalanced. Taken together, these limitations slightly decrease the trustworthiness of the study's findings.

34)

Šturm, P., Przedlacka, J., & Rojczyk, A. (2022). Phonetic imitation of t-glotalling by Czech speakers of English. *Linguistica Pragensia, 1*, 142-165.

Šturm et al. (2022) conducted a study to investigate Czech speakers' ability to imitate native English realisations of the phoneme /t/ as [ʔ] in shadowing tasks.

They hypothesised that, due to patterns of glottalization in Czech, participants would imitate [ʔ] better in intervocalic contexts than in non-prevocalic contexts. The alternative hypothesis was that, due to frequency of the learners' English input, the opposite pattern would be true.

Participants were 30 native speakers of Czech, 15 males and 15 females. Their age ranged from 19 to 39 years old, with a mean of 25.3. Most participants studied at the Faculty of Arts, Charles University, Prague, and a minority worked at the university library or had no connection to the university. There were three different relevant profiles of students: students of English studies (i.e, expert English users, $n = 2$), students of phonetics as a full programme (i.e., expert listeners, $n = 5$) and other participants ($n = 23$). Results from the first two groups were analysed separately, given their expert knowledge. Participants also had different levels of English proficiency (1 A level, 16 B level, 13 C level), which were merged into intermediate (A and B) or advanced (C) for analysis.

Words selected for the experiment involved 16 words with an intervocalic /t/ (VtV), 16 words with a non-vocalic /t/ (in VtC or Vt#) position, as well as 16 filler words without a /t/ or /k/ segment. There were also nine words with a /t/ segment in the same positions that were not shadowed, and nine words with a /k/ segment in a non-prevocalic position, to test whether participants generalised beyond the shadowed words. The recordings of words came from a corpus of Southern British English, in which speakers had taken part in informal interviews, and YouTube videos. The accents of the speakers were Estuary English and Cockney.

Participants completed a pre-test (baseline) reading task in which they read words appearing on a screen. They then completed an imitation task in which they heard the recording of the words and repeated it. Finally, they completed a post-test reading task (identical to the baseline test). The three blocks were presented in the order described, but the order of items within them was randomised.

Recordings from these blocks produced a total of 3.960 observations (30 speakers*132 items). Each token was examined in Praat and analysed for phonetic realisation of the target segments via a combination of spectrographic and waveform cues, and auditory analysis. The tokens were then categorised as glottalized or non-glottalised. The variable of glottalization was analysed using mixed-effects logistic regression in R. The regression models included fixed effects (task, position, sex, expertness, level) and random effects (item and participant), with biological sex, expertness, and proficiency included as control variables.

Results suggested that very few tokens were glottalized in the baseline reading task, with any cases being in the non-prevocalic contexts. There was a much larger proportion of glottalized tokens in the shadowing task, with somewhat higher rates in intervocalic positions. In the post-reading task, there were lower rates of glottalization. The logistic model indicated that task was a significant predictor of glottalization ($p < 0.001$) but position was not ($p = 0.889$). Post-hoc pairwise comparisons indicated that the non-prevocalic position in the post-test was associated with significantly higher rates of glottalization as compared to VtV ($p = 0.0013$), unlike in the shadowing task where there was no significant difference ($p = 0.586$). Further comparisons indicated that the shadowing task was significantly different to other tasks ($p < 0.001$) and that the post-test had significantly higher rates of glottalization than the pre-test ($p = 0.018$).

In a logistic model for the shadowing task, expertness was statistically significant ($p = 0.040$), meaning experts had a higher probability of glottalization. There was also a significant interaction of position and expertness ($p = 0.034$), meaning expert participants glottalized more in the intervocalic position than the non-prevocalic position. No significant interactions emerged in the logistic model of the post-test.

Non-shadowed items did not differ from shadowed items at pre-test, but at post-test there was a descending order of glottalization rates from shadowed to non-shadowed words and non-shadowed /k/. However, these differences were not substantial.

From these results, the authors conclude that participants appeared quite responsive to glottalization in immediate imitation, but did not retain these productions at post-test. This suggests that imitation, rather than learning, was responsible for the increase during shadowing. They also note that some degree of generalisation of glottalization must have occurred, as participants increased in their glottalization of non-shadowed and shadowed words alike. However, they caution that the number of glottalized tokens at post-test was generally low. They also conclude that expert listening skills, knowledge of the L2, and the target sound appearing in the same position (VtV) as the target and source language, seemed to facilitate imitation.

Overall, this study seems robustly designed and trustworthy. However, as shadowing does not fit the definition of this review, and is simple imitation, the design is not considered appropriate and results are not considered relevant. For this reason, the study is not discussed in the results section of this dissertation.

35)

Sumiyoshi, H. (2014). ***Exploring the effects of the shadowing method: case studies of Japanese language learners at an Australian university*** **[Unpublished doctoral dissertation]. Macquarie University.**

Sumiyoshi (2014) conducted a series of studies on shadowing as part of her PhD research. All studies were conducted at an Australian university with university-level students of Japanese over four semesters, from 2016 to 2017. Students were enrolled in a range of Japanese courses, all taught by the author. Some studies were related to listening skills, and others were related to pronunciation skills and student perceptions. Only the latter two types of studies are reported on here, as they were the only deemed relevant to answering the questions of this systematic review.

Study 1)

Study 1 investigated improvements in pitch accent accuracy after shadowing training, and student perceptions of the intervention.

The study involved 20 students enrolled in an intermediate level spoken Japanese course. Participants were six males and 14 females, most of whom were between the ages of 18-24 ($n = 18$), with others between 24-30 ($n = 2$). Most participants were majors in Japanese ($n = 16$), with other participants studying related majors like International Studies ($n = 1$), Translation ($n = 1$) or Linguistics ($n = 1$). It should be noted that no information is given on participants' L1 in this study.

The study involved students undertaking shadowing homework over a 7-week period (although no homework was undertaken in weeks 5 and 8 as there were speech and interview tests). Shadowing texts were selected from the book *Japanese Pronunciation Practice through Shadowing*. Texts were seven monologic audio recordings of between 47 and 64 seconds. The materials were ordered so that the speed would increase gradually each week. Their readability was assessed as between upper elementary and upper intermediate level. Students shadowed one text per week and had five days to practice shadowing before submitting a final recording. The teacher gave them feedback on their mistakes and total accuracy was calculated out of 100%. The students also practised whole class shadowing after every assignment with a focus on common mistakes.

Data was collected through recordings of participants' shadowing, with recordings from week 2 and week 10 of the semester analysed. A survey was also conducted to explore student perceptions, with sections related to perceptions of improvement in 1) listening 2) speaking 3) fluency 4) pitch-accent 5) pronunciation 6) native-like speech 7) speed variation 8) simultaneous listening and speaking 9) intrinsic motivation 10) comparison to other listening/speaking practice 11) intention to continue, as well as some open-ended questions on overall perceptions of the intervention.

The model audios from weeks 2 and 10 were analysed for pitch accent by an online prosodic reading tutor system, to maintain consistency and objectivity. 42 pitch accent falls were detected for the week 2 audio, and 41 for the week 10 audio. Praat was then used to project participants' shadowing audio in an F0 curve to help the rater, who appears to have been Sumiyoshi, compare pitch falls to the model audio. It should be noted that it is unclear whether the rater was blind as to whether recordings were from week 2 or week 10. This is important as knowledge of this fact could have introduced an element of bias into the results. There also appears to have been only one rater, although this is mitigated somewhat by use of the online system. Mean and median scores, and standard deviations, were calculated for the survey results. Statistical analyses were used to investigate whole group change, and scatter diagrams were used to investigate individual changes. Open responses were categorised, and frequency counts were produced.

Pitch accent accuracy in shadowing performance showed improvement from week 2 to week 10, with mean scores increasing from 81.48 to 89.56 and median scores increasing from 82.5 to 91.37. A Wilcoxon signed-rank test indicated that median scores increased significantly over time, and Cohen's d was estimated at 0.57, indicating a median effect size based on Cohen's (1992) guidelines. The scatter diagram indicated that the majority of participants (n = 16) showed positive growth in pitch accent accuracy, whilst four participants showed negative growth. Survey results showed that participants all agreed that shadowing was effective for improving listening skills (M= 5.28), with over 80% also agreeing that it was effective for pronunciation, fluency, and speed. Participants were less positive about shadowing being effective for improving pitch accent and native-speech, but were still positive overall (78.3% positive, and 61.7% positive, respectively). Participants who showed negative trends in pitch accent responded less positively about pitch accent than the mean score, and also responded less positively for overall questionnaire items than the mean score. Participants who improved, however, responded more positively than the mean score for all questionnaire items. 19 participants added comments about the positive aspects of shadowing, and 15 added comments about the negative aspects. In addition, 43 key comments were positive, whilst 20 were negative. The most frequently mentioned positive aspect was improvement of pronunciation

(n = 10), followed by listening (n = 7). The most frequently mentioned negative comment was difficulties with the overly fast speed of the materials (n = 8). Some comments related to speed, accuracy, pitch-accent and fluency were mentioned as both positive and negative aspects, but did not overlap in comments from the same participants.

From these results, the author concludes that participants developed more accurate linkage between the perception and production of high-low pitch accent over the course of the study, and that there was a close relationship between degree of improvement in pitch-accent accuracy and participants' perceived attitudes towards shadowing. She also notes that participants held different opinions of what aspects of shadowing were positive or negative, meaning that there can be individual differences in responses to the intervention. This study should be highlighted for its detailed descriptions of the shadowing procedures used, the detailed analysis of the difficulty of each shadowing text, and some measures taken to maintain objectivity and reliability in measures of pitch accent. However, it should be noted that the issue of whether the rater of pitch accent was blind with regard to the week of the recording is an important one, and may limit the trustworthiness of the findings, particularly because there was only one rater. There is also no information given about how and by whom the surveys were conducted. If they were distributed/conducted by the teacher (which appears likely given that Sumiyoshi was both teacher and researcher in the studies) and if students were not ensured that their responses were anonymous, this could have meant students presented overly positive views of the intervention. Finally, the use of shadowing recordings may not have been the most ecologically valid measure of pronunciation: use of spontaneous tasks would have improved this issue.

Study 2)

Study 2 was a quasi-experimental study comparing improvements in pitch accent between a shadowing and non-shadowing group.

Participants were 46 students of Japanese at an Australian university enrolled in either Intermediate Spoken Japanese of Intermediate Japanese I. Participants were 26 females and 20 males, although it should be noted that no information is given about their age, L1, or majors.

Participants were divided into a control (n = 34) and experimental group (n = 12). It is unclear how participants were divided, but appears to have been guided by university policy on which courses students were enrolled in and policies about assessment tasks connected to these courses. For this reason, the experimental group were students enrolled in both Intermediate Japanese I and Intermediate Spoken Japanese, and the control group were students just enrolled in Intermediate Japanese I. This resulted in a number of other differences between the control and experimental groups, such as overall numbers, gender balance, and teaching input hours.

From the limited description available, it appears that the students in the experimental group received an additional 2 hours of instruction per week, and the shadowing homework as described in study 1. It is unclear what the control group were doing in their classes.

Both the control and the experimental group took part in 5 recitation homework tasks over the same semester and the shadowing homework. These tasks were all assessed for number of morae, speed, readability, difficulty, number of sentences, and words per sentence. They were all of an intermediate level. A model audio recording of each recitation was given to students at the beginning of each week, and students recorded themselves reading the passage at the end of the week.

Recordings from weeks 2 and 10 were analysed for pitch-accent accuracy. To address issues of bias and objectivity, participants' IDs were replaced with numerical IDs and random numbers were given in W2 and W10. This meant that the marker was blind as to whether participants were in the experimental or control group, and there was no way to identify which week the recordings were from. The recitation audio files were analysed for pitch accent accuracy by using the online prosodic reading system to detect pitch accent falls in the model audio. The marker then appears to have rated students' recordings as in study one, although this information is unclear.

Results indicated that both the experimental and control groups improved in pitch accuracy from week 2 to week 10, with mean scores increasing from 76.75 to 79.56, and from 70.12 to 71.6, respectively. Median scores in the experimental group and the control group were significantly higher at post-test than pre-tests ($p = 0.041$ and $p = 0.295$, respectively). However, the effect size for the experimental group was medium ($d = 0.64$) and the effect size for the control group ($d = 0.15$) was small. In the experimental group, eight participants showed positive growth (66.7%) and four showed negative growth (33.3%). In the control group, 21 participants showed positive growth (61.8%) and 13 showed negative growth (38.2%). There was a positive correlation in the W2-W10 improvements in the shadowing and recitation tasks, but this was not statistically significant.

The author concludes that the experimental group showed stronger improvement in pitch-accent than the control group, indicating that shadowing had an effect on pitch-accent accuracy in recitation tasks. These results should be taken with caution, due to the fact that there was only one rater, the small differences in scores and levels of statistical significance between the control and experimental group and, most importantly, the important differences in the composition of these groups. The fact that the experimental group were enrolled in Intermediate Spoken Japanese may have meant they were much more motivated to practise speaking, as the author notes, which may have led to increased practice and therefore higher scores. The additional study hours they received may also have contributed to their improvement. It should also be noted that no information is given on participants' L1 in each group. Taken together, these confounding variables could explain the small differences in performance between the two groups, making it very difficult to gauge the effect of shadowing homework.

Study 3)

Study 3 was designed to investigate participants' perceptions of shadowing, via a questionnaire administered after a period of studying shadowing.

It involved 36 participants at the same Australian university as the other two studies. Participants were 25 females and 11 males, all between 18-24. Most were majoring in Japanese ($n = 28$), and all were enrolled in an advanced Japanese language unit. Many spoke English as an L1 ($n = 21$), although others spoke Chinese ($n = 11$) , Korean ($n = 2$) or other languages ($n = 2$). One third of the participants had experienced shadowing in the previous semester.

Limited information is given on what the shadowing intervention involved, for instance the duration of the study. What does appear clear is that students undertook shadowing as a homework activity and that model audios were recorded by the teaching instructor, who was a

native Japanese speaker. It appears that students had different shadowing material each week, and that they had to repeat it at least six times before recording their shadowing performance.

After their shadowing practice period concluded, students each completed a questionnaire. The questionnaire was comprised of three sections: demographic information, questionnaire items, and open-ended questions. Questionnaire items fell into five categories: perception of improvement in listening and speaking, attitude towards shadowing performance and satisfaction, intrinsic motivation and extrinsic motivation, and cost. Motivation items were based on Pintrich et al.'s (1991) Motivation Strategies for Learning Questionnaire, items related to cost were developed from Hamada's (2011) study, and other items were constructed with reference to the shadowing construct itself. Cronbach's alpha was calculated for internal consistency and reliability of each category, with improvement (.907) and attitude (.848) high, and motivation (.663) and cost (.671) lower. The items used a 6-point Likert scale, in which 6 was "strongly agree" and 1 "strongly disagree". The items were randomly shuffled, and rearranged manually where questions in the same category appeared in succession.

To analyse survey data, frequencies and percentages of participants' responses to questionnaires were calculated, and an exploratory factor analysis was performed. Participants were also divided into high performance (n = 18) and low performance (n = 18) groups, to investigate differences in attitudes to shadowing depending on performance skills. Open responses to the survey were analysed qualitatively.

Results indicated that the majority of participants perceived improvement in speaking and listening due to shadowing (over 80% agreed with the positive aspects of shadowing in these areas). All participants agreed that feedback is very useful for shadowing, and 97% believed they could become better with practice. 36% did not believe it was necessary to shadow at a faster speed than they could speak. 64% of participants showed an intention to continue shadowing after the study finished. Responses related to items on motivation were mixed, with some positive (e.g. "correcting mistakes after feedback is important"), and some mixed. In terms of cost, respondents tended to indicate that they procrastinate in completing shadowing tasks and feel frustrated about the fast speed of the audios.

Factor analysis indicated four factors as the most interpretable solution from results, which were named "positive speaking", "accurate listening", "good appearance" and "feedback for marks", based on the items loaded on to them.

The most salient trend in relation to the high-performance group was the fact that intention to continue shadowing was correlated with the challenging nature of the task, which the author connects to higher performing students' ability to set manageable goals and realistic ideal selves. Intention to continue shadowing amongst the lower performance group, however, was mostly related to extrinsic motivation: perceived improvements in bottom-up processing skills.

In terms of responses to open-ended questions, 97% of participants mentioned positive aspects of shadowing and 89% mentioned negative aspects. Overall, there were more comments about positive aspects (n = 72) than negative aspects (n = 40) of the intervention. Both performance level groups made the same number of positive comments (n = 36), but there were more comments from the lower performance group (n = 22) compared to the higher-level group (n = 18), particularly with regard to shadowing speech (12 compared to 6). Other differences between groups included more comments about improvements in speaking and listening in the LP group compared to the HP group, and more comments about improvement in pronunciation

and native-like speech amongst the HP group. In terms of suggested improvements to shadowing, the most common type of comment was allowing for resubmission after feedback to observe improvement in the same material (n = 3) or adapting content (n = 4), either by having English translations available or improving the interest of the materials.

From these results, the author concludes that the majority of participants perceived shadowing as effective for developing speaking and listening skills, with all agreeing on the importance of feedback. However, individual differences were found in attitudes, particularly with regard to the speed of materials, with some participants finding the challenge of fast speed audios motivating and others perceiving it negatively. To deal with this issue, the author recommends creating two or three versions of shadowing materials at different speeds to meet different shadowing proficiencies.

As the author acknowledges, the data presented are limited by the small sample size, particularly with reference to the factor analysis. Another key limitation of this study is the lack of description of the duration of the shadowing intervention, making it difficult to compare it to other studies. In addition, it is unclear if the teacher conducted the survey in class, and whether students were informed that their responses were anonymous, which could have led students to present overly positive views of the intervention. Finally, exploring perceptions and attitudes via a survey could have been greatly enriched by other sources of qualitative data, such as semi-structured interviews or classroom observations.

Overall, based on the results of the three studies, Sumiyoshi concludes that shadowing is an effective way to improve speaking skills, though she cautions that it places a heavy cognitive load onto learners and, as such, should be practised in brief sessions as one technique amongst many for practising speaking skills. The data do seem to support her claim as to the effectiveness of shadowing, although the results of study 1 and study 3 are much more robust than study 2. However, it should be noted that, overall, the studies provide a very narrow measurement of "speaking skills", investigating only pitch accent through either shadowing or recitation (read aloud) tasks. To gain a fuller picture of the impact of shadowing on pronunciation, other measurements, such as impressionistic judgements, or measurements of duration and intensity, could have been used. In addition, it would have been beneficial to include measures of spontaneous speaking, to gauge the extent to which participants' pronunciation improved in more natural speaking.

36)

Sumiyoshi, H., & Svetanant, C. (2017). Motivation and attitude towards shadowing: learners' perspectives in Japanese as a foreign language. ***Asian-Pacific Journal of Second and Foreign Language Education, 2*****(16).** https://doi.org/10.1186/s40862-017-0039-6

Sumiyoshi and Svetanant (2017) conducted a shadowing intervention with 36 advanced learners of Japanese in an Australian University. Participants were aged 18-24, and were predominantly L1 English speakers (Chinese n = 11, Korean n = 2, other n =2). After the intervention, the authors conducted a survey, including three Likert-scale questions and three open-ended questions. The questionnaire aimed to explore four main areas of investigation. Firstly, whether participants considered shadowing effective. Secondly, what they considered the positive and negative features of the intervention to be. Thirdly, if there was any difference in participants' intention to continue shadowing in the future and their performance and

motivation during the intervention. And, finally, the factors that encourage participants to continue shadowing in the future.

Frequencies and percentages calculated for different items showed that the majority of participants considered shadowing to be effective for developing speaking and listening skills, with 64% stating an intention to continue shadowing after the study. In addition, 100% of participants stated that feedback on their shadowing was important. However, opinions were not entirely positive, and views on shadowing appeared complex. For example, some participants indicated that they tended to procrastinate in completing their shadowing assessments, and other expressed frustration about the fast speed of the audio track.

Analysis of open responses further highlighted the mixed attitudes toward shadowing, with a large number of comments on both the positive and negative aspects of the intervention. In total, there were almost twice as many comments for positive aspects ($n = 72$) than for negative aspects ($n = 40$). There were equal number of positive comments in the high proficiency (HP) and low proficiency (LP) groups, but more negative comments from the LP group ($n = 22$) compared to the HP group ($n = 18$). In addition, the LP and HP groups had different perspectives on which features of shadowing were positive, with the LP group more frequently highlighting improvements in speaking ($n = 11$) and listening ($n = 11$), and the HP group more frequently commenting on improvements in pronunciation ($n = 7$) and native-like speech ($n = 7$). The authors suggest that the LP group's focus on speaking and listening may be due to the fact that their lower level left more room for improvement in these skills, more related to improving basic bottom-up processing. HPs' focus, on the other hand, on more native-like speech and pronunciation suggested more awareness of specific, more subtle elements of speaking skills.

The most salient difference between the two groups in terms of negative aspects of shadowing was the number of comments against the shadowing speed by LP participants ($n = 12$) compared to HP participants ($n =6$). Whilst the authors acknowledge that difficulty with speed at lower levels is normal, they suggest that in future interventions two or three different speeds should be provided to accommodate different levels.

When asked about how to improve shadowing, participants' most common comments were about the frequency, increasing the total number of weeks, re-submission after feedback to observe improvement in the same material, or more shadowing opportunities to reduce the pressure per submission. Two participants also suggested having English translations of the material available, and two wanted more interesting materials.

To explore the link between motivation and intention to continue shadowing, Pearson's correlation coefficients were calculated using the item "continue shadowing after completing this unity" as the dependent variable, and other items as the independent variables. The most salient trend in the HP group was the correlation between intention to continue and like shadowing because it is a challenging activity, which the authors link to a realistic ideal self and integrated regulation amongst these participants. The LP group showed a different trend: intention to continue was correlated with different items such as perceived improvement in shadowing performance and pronunciation and speaking skills, as well as valuable experience and challenging. The authors link this to identification in motivation theory (willing to do something because it is relevant).

Whilst overall, student perceptions of the intervention seem to have been explored robustly, the links to motivation are weaker. Firstly, Cronbach's alpha test suggested that the motivational items on the survey had lower internal reliability than other blocks. Secondly, there are some weaknesses at the level of matching interpretation of results to motivational theory. It seems reasonable to connect LPs intention to continue shadowing with the idea of identification in Ryan and Deci's (2000) model, as, taken together, the items in question (Sf1:shadowing is a valuable learning experience, Pf1: become better at shadowing after 1 week, Pf4: can become better at shadowing if practice more, Sp1: shadowing is effective in improving pronunciation, Sp4: pronunciation became better after shadowing, Sp3: speaking skills improve the more practice shadowing) are directly related to doing an activity because it is relevant and improves skills. However, making a link between ideal self theories and the correlation between HPs intention to continue and the item IM1 (like shadowing because it is challenging) does not seem warranted, as the items are not directly related to the theories of ideal selves.

Additional reporting and methodological weaknesses also limit the findings of this study. Firstly, analysis of open responses forms a significant part of the paper. However, the type of analysis applied is not specified and only listed as "qualitative". Secondly, exploring connections between motivation and intent to continue shadowing through statistical analysis does not seem entirely appropriate, particularly given the mismatches between theory and results described above, and the small sample size of 36. With such a sample size and the topic at hand, semi-structured interviews would have provided much richer data.

In summary, the authors explored student perceptions and the positive and negative factors as well as could be done through mainly quantitative research. However, exploration of motivation and the reasons for which participants wanted to continue shadowing in the future was more problematic, due to low internal reliability of items, poor connections between items and some elements of motivational theory, and the study design of quantitative statistical and factor analyses for a small sample. Richer qualitative data could have been collected through interviews and classroom observations, for example.

37)

Teeter, J. (2017). Improving motivation to learn English in Japan with a self-study shadowing application. *Languages, 2*(19). https://doi.org/10.3390/languages2040019

Teeter (2017) conducted a study to explore student motivation and attitudes related to shadowing practice.

Participants were 1001 first-year students of English at a university in Japan. All students were enrolled in 30 intact academic English classes, but most were science and engineering majors. Students were aged 18-23, with 72% males and 28% females. 97% of the sample identified as Japanese. Four different teachers taught the classes. Student proficiency was not measured directly, but the average English proficiency at the university was reported to be 78 out of 120 on the TOEFL-ITP test, which indicates an intermediate proficiency level.

The study involved 12 weeks of shadowing practice, assigned to students as homework by their teachers in a specially designed shadowing app. Students shadowed five assigned texts per week. Texts were around 1 minute long, and sourced from materials designed to prepare students for the TOEIC test. Students had flexibility in how they shadowed, and were able to

choose whether or not to do pre- and post-shadowing, whether to shadow with or without a script, and how many times to practise before submitting their recordings to their teachers.

To explore student attitudes to the intervention, Teeter conducted a pre- and post-intervention survey. A survey involved a 47-item questionnaire (in Japanese) administered before the intervention and after 14 weeks. Items other than demographic information were all six-point Likert scale questions with no neutral option. The questionnaire was developed from other previously tested and administered questionnaires on the motivational self system. The different modules explored different aspects of the L2 motivational self system: instrumental, linguistic self-confidence, anxiety, attitudes towards listening comprehension practice in the L2, interest in oral communication in the L2, ideal L2 self. There were also questions on how often students met speakers of English, heard English spoken in the media, and listened to music in English. Seven additional items were added to the post-intervention version to explore students' enjoyment of shadowing practice, perception sof its efficacy for improving listening comprehension, and intention to use shadowing in the future. There was also a question about how much students used shadowing per week. Some students without access to technology at home completed the questionnaire in class, whilst others completed it outside of class.

Before analysis, responses from students who did not identify as Japanese (n = 18) and cases of missing data were removed, resulting in a sample of 987 responses for the pre-test and 747 responses for the post-test. A range of statistical analyses were then performed on SPPS, including: independent samples t-tests to examine whether there were significant changes over time in motivation and attitudes; independent samples t-tests on constructs of motivation and attitudes; independent samples t-tests on perceptions of English speaking and listening proficiency; cross tabulation to determine how contact with English changed during research periods; Spearman's correlations on all items; descriptive statistics on additional items related to shadowing and reliability analysis with cross tabulation.

Significant increases were found for the following items when comparing pre- and post-test results: "English listening comprehension is easy", "I am good at learning English", "I know effective ways to study English listening comprehension", "I get excited when my teacher speaks to the class in English", "I believe I will be able to understand spoken English with little difficulty if I keep practising listening comprehension", "I can imagine myself speaking English to other English speakers like a fluent speaker in 2 years". In addition, there was a significant decrease in the items: "I feel like other students in my class are better than me at English", "I really want to be able to understand spoken English", "Being able to understand spoken English is important to me". Analyses also indicated that, overall, there were significantly higher means at post-test for items associated with the instrumental construct and linguistic self-confidence construct. Self-rated proficiency also significantly improved over time. Self-reported frequency of shadowing practice was correlated with all instrumental questions (weak to moderate), some items related to interest in communication (weak to moderate), most anxiety items (very weak to weak), most ideal self items (weak to moderate), and some attitude to listening comprehension items (weak to moderate). Importantly, students who did less shadowing were less likely to report improved listening comprehension due to shadowing. The median response to the item "shadowing is interesting" was 3 (close to slightly disagree), and the medians for "shadowing is boring", "shadowing is an effective way of learning English" and "I want to use shadowing to practise listening comprehension in the future" were 4 (close to slightly agree). Reliability analyses indicated that between 8.76% and 29.09% of respondents may not have been completing the survey in a consistent way, for example with respect to the items "shadowing is boring" and "shadowing is interesting".

From these results, the authors draw several robust conclusions: that students' linguistic self-confidence, ideal L2 self-concept (in relation to English listening comprehension and speaking) and self-rated listening proficiency improved significantly after the shadowing intervention. They also note that anxiety, at least with respect to comparison with other students, may have decreased through shadowing practice.

However, some of their other conclusions are less well supported by the data. Firstly, and most importantly, they note that students viewed shadowing favourably. However, the median response to the item "shadowing is interesting" was close to "slightly disagree" and the median response to "shadowing is boring" was close to "slightly agree". In combination with the loss of participants throughout the study, probably those not interested in shadowing, suggests that student perceptions of the intervention were not as positive as the authors claim. In addition, the authors claim that shadowing may have had an impact on students' interest in communicating in English, due to positive correlations between practice and interest in communication. It should be noted, however, that interest in communication actually decreased during the study. The authors also make a number of claims about correlations between frequency of shadowing practice and particular items, for instance that students who spent more time shadowing had more positive views of their ideal self, more motivation and desire to use shadowing in the future, and self-rated their listening comprehension more highly. However, some of the correlations they refer to here are weak, and at most moderate, so should not be over-interpreted.

Overall, the study's strengths are its detailed description of the shadowing app and practice materials, and its large sample size. The authors also controlled for the variable of L1 by removing data from non-Japanese participants. Significant weaknesses of the study include the lack of a control group and the fact that the four different teachers involved in the research could have influenced student motivation differently. In addition, the fact that some students completed the survey in class and others outside of class could have influenced the results, as could the fact that survey responses may not have been anonymous (this was not specified). Finally, the authors do not explain their lack of post-test data in detail. It could have involved a drop out of 236 students (as there were only 747 responses to the post-test), which should be considered in discussion of the results. This is particularly important as, if students dropped out because they did not like shadowing, the results presented on attitudes and motivation could be overly positive due to a self-selection bias amongst the participants. Finally, there is the issue of self-reporting of listening proficiency, which may not be a reliable measure.

Overall, whilst some of the results of this study seem sound, others should be taken with caution.

38)

Wang, X. (2018). The study of shadowing exercise on improving oral English ability for non-English major college students. ***Advances in Social Science, Education and Humanities Research, 120*****, 195-200.**

Wang (2018) conducted a longitudinal study exploring the impact of 10 weeks of shadowing practice on students' oral English ability and attitudes to oral English.

Participants were 40 sophomore Chinese university students studying but not majoring in English. 28 males and 12 females participated in the experiment. Wang observes that there were "no obvious differences in English proficiency" between the groups, but their exact proficiency and L1 are not provided. Participants were divided randomly and equally into a control and experimental group.

Both groups studied for a total of 10 weeks, with one hour of class per week. The learning materials, topics, and schedule of the classes were the same. Both classes were taught by the author of the study. Classes tended to focus on listening to news, stories, or dialogues, and discussions of topics to make a speech. However, the difference between the two groups was that the experimental group integrated shadowing into the classes and had more of a focus on oral expression, whilst the control group did not. All 40 students were required to review and practice after class, but the experimental group also had to record their after-class shadowing.

Shadowing materials were selected from *The Intermediate Course of Interpretation* book and VOA Special English. The difficulty of shadowing increased as the training progressed, moving from simple sentences to difficult and long sentences, then paragraphs. And from 1-2 minutes to 2-3 minutes of shadowing time.

To collect data, a pre- and post- questionnaire, with items about attitudes to oral English (interest, recognition of importance, frequency of practice, time spent on it), and understanding of shadowing was administered. Questions were all 5-point Likert scale questions. A pre- and post-test of oral English was also administered (from *New Horizon College English - Viewing-Listening-Speaking*). The post-test involved making an impromptu speech about environmental problems, without advanced warning. However, it is unclear whether the pre-test was similar or different, which is important information as it could have influenced results. It should be noted that no information is given about the validity and reliability of any of the tests used.

To analyse data, recordings of the pre- and post-oral tests were given fluency scores, which was calculated as the number of words spoken per minute. These scores were then compared using SPSS and t-tests. Frequency counts and percentages were calculated for survey data.

The pre-test questionnaire indicated that, before the experiment, 31 students (77.5%) thought oral English was very important and 21 (52.5%) were not interested in oral English. 6 students (15%) had consciously practised oral English, but for no longer than 30 minutes, and 35 (87.8%) students had never heard of shadowing. The post-intervention questionnaire indicated that 18 (90%) of students thought that the shadowing exercise was interesting, with 19 (95%) reporting that their oral English fluency had greatly improved. All students (20) believed that their oral English ability had improved, and 17 (85%) believed shadowing improved their confidence and interest in speaking. 17 (85%) students stated that they would continue to practise shadowing in the future. In terms of oral ability, t-tests indicated that there were no significant differences in fluency between the groups at pre-test. At post-test, the experimental group scored significantly higher than the control group (71.8 vs 57.4, respectively, $p = .001$).

From these results, Wang (2018) concludes that shadowing can greatly improve students' oral ability and their interest and confidence in oral English. Whilst the results of the study are clearly positive, findings should not be over-interpreted for several reasons. Firstly, if the students carried out the questionnaire in class with their teacher (who was also the researcher) and were not informed that their responses were anonymous, the perceptions presented may be overly positive. Secondly, the author should be clear that, whilst students self-reported

improvement in overall oral ability, this self-report may not be the most valid or reliable measure of improvement. In addition, the only objective measure of oral ability explored in the study is fluency, as measured by words spoken per minute. It is also unclear what the pre-test of oral ability involved, so difficult to assess the extent of a possible repetition effect on fluency gains. Because of this, the study should be considered to provide initial support for an increase in fluency with shadowing, and potential gains in other aspects of oral ability, as reported by students. The study could have been improved by using more measures of overall oral improvement, for example acoustic measures of pitch, duration, and intensity or impressionistic judgements, and additional qualitative data to supplement survey findings (e.g. open questions in the survey, semi-structured interviews, classroom observations).

39)

Willardson, V. (2014). ***The effectiveness of computer-enhanced shadowing and tracking pronunciation exercises for intermediate level foreign language learners*** **[Unpublished master's dissertation]. Brigham Young University.**

Willardson (2014) conducted a study to explore whether computer-assisted shadowing and tracking could help participants improve French pronunciation and become more aware of their own skill level.

Participants were 19 high school students of French at a school in the state of Utah, USA. 12 were male, seven were female, and all but one, who was a native speaker of Chinese, were native speakers of English.

During the 10-week intervention, students completed 5-10 minute tracking and shadowing activities as a class three times per week, using conversations from the textbook *D'accord!*. They also completed one 20-30 minute individual session in the language lab, shadowing videos from the *Ma France* program produced by the BBC.

Identical pre- and post-tests were used to measure pronunciation improvement. These tests consisted of one free-response picture description task, in which students spoke for 1 minute about the people and activities in a picture, and one read-aloud task, in which students read from a text containing six short dialogues. Tests were scored by three expert raters, who were blind as to whether were analysing pre- or post-test recordings. Raters graded speech samples on a Likert scale in which 1 represented heavily accented speech difficult to understand, and 5 near-native pronunciation able to communicate with all. They were instructed to base scores on general accent, word stress, sentence stress, and overall intonation, and were instructed to attend to sounds like the French R and different French/English vowel sounds as precise markers at word level. They were also instructed to ignore aspects of speech like fluency and grammar.

Six surveys were also conducted to explore student use of materials in the lab sessions. Students rated how often they looked at subtitles (and in which language), how often they needed to stop the video, how difficult it was to say what the speaker said, the speed of the video, whether it was too fast, how important it was for them to understand what was happening in order to repeat. They also separately rated the in-class exercises and provided comments about why they liked/disliked these activities, what helped them, whether they were of an appropriate level. Students also gave final comments after the study on their overall experience.

In addition, the researcher made observations throughout the course of the program to record her perceptions of how the study was progressing.

Teacher observations seemed to suggest a lessening reliance on captions throughout the study, and student apprehension when dialogues were too fast. Students seemed positive about the intervention, although some found filling out the survey repetitive. Observations also detailed the teacher's enthusiasm about the intervention.

Qualitative survey results showed positive responses to the program over time. In week 1, less students shared positive comments (26%) about the intervention and more had mixed (26%) or negative feelings (37%), often because they felt overwhelmed. However, at the end of the program, more students seemed positive about the intervention (74%). Themes identified to explain this positivity included: increased interest in French people and culture, the autonomy of the lab exercises, the availability of the subtitles, and progression in language skills. In week 1, most students (18/19) believed that the features of the intervention that helped pronunciation were: hearing correct pronunciation, reading the captions, and speaking along with the videos. One student felt that the in-class activity was too chaotic, and helpful as they couldn't hear clearly. At the end of the program, all participants expressed positive comments, believing that the intervention helped pronunciation because of: listening to native speakers, reading along with the subtitles, pronouncing out loud, and having the autonomy to go at their own pace in the lab exercises. In additional comments on their perceptions of the program, students were mostly positive (but some very apprehensive) in week 1, with all students giving positive comments after the intervention.

Quantitative survey results provided insights into student habits in the lab sessions. On average, students used subtitles extensively (average score of 3- often). On average, students did not stop the videos much at the beginning (average score of 2, rarely), but there was an increase in week 7. On average, students considered that it was either not difficult or somewhat difficult to repeat what the speaker said, with no changes over time. On average, students found the speed of the video clips average or fast but manageable. On average, students did not find it important to understand what speakers had said in order to repeat. On average, students found the lab work helpful, and their scores on this question statistically improved ($p<.016$) between the first and last survey. On average, students were less satisfied with the in-class exercises, with average scores of 3 (1- not helpful, 5- very helpful) that did not change over time.

In terms of pronunciation improvement, average scores on the spontaneous measure increased from 7.2 to 7.6 from pre- to post-test, however this was not significant ($p = 0.20$). When students were divided into top, middle, and bottom groups, there were some significant differences. The top group's scores improved from 9.3 to 10.5, and was significant ($p = 0.05$). The middle group's scores decreased from 7.8 to 7.1, but was not significant ($p =0.16$). The bottom group's scores increased from 3.5 to 4.4, and was significant ($p = 0.03$).

Read aloud scores, however, showed a statistically significant improvement ($p = 0.00005$) for all groups from pre- to post-test, from 7.5 to 8.6. The top group showed an improvement of 9.8 to 11.4 ($p = 0.02$), the middle group showed an increase from 7.5 to 8.4 ($p = 0.04$), and the bottom group improved from 4.3 to 5.6 ($p = 0.005$). In both measures of pronunciation improvement, improvement in the bottom group was the largest and most significant.

From these results, Willardson (2014) concludes that shadowing and/or tracking exercises were effective in improving the accuracy of students' pronunciation, particularly reading aloud skills,

and that students responded positively to the intervention. She advises teachers interested in implementing such a training program to help motivate students by: letting them make their own selection of viewing material and giving them a sense of control over what they watch, using videos with native actors in real situations, making sure the difficulty level is appropriate, using captions, preparing students for in-class exercises by asking them to speak softly.

Overall, this is a highly relevant study for this systematic review. Its strengths include the detailed description of the procedure followed, the use of data triangulation to explore students' attitudes (combining classroom observations and surveys), the use of a controlled and spontaneous measure of pronunciation, and the reduction of bias in rating by use of three raters and blind rating with regard to pre- and post-test recordings. It should be noted that this last point could have been improved even further through reporting of inter-rater reliability (these calculations are given in Martinsen et al. (2017)).

However, there are some weaknesses of the study that should be acknowledged. Firstly, it is difficult to untangle when students were "shadowing" and when they were "tracking", as the definitions given overlap and students seemed to have the option to do both in all activities. The author also does not define shadowing or tracking with any reference to the literature, defining tracking as speaking along with an audio and shadowing as having the ability to pause. However, as both the activities referred to as "tracking" and the activities referred to as "shadowing" both fit under the umbrella of shadowing, as defined in this review, this is not as problematic as it appears. One issue that should be noted, however, is that some students appeared to be reading ahead of the speakers in the videos by using the subtitles. Whilst the researcher corrected this behaviour, it appeared frequent. Lack of control over exactly what participants were doing in the lab again does make it difficult to be completely certain that all participants were indeed shadowing.

Secondly, minimal information is given about the pre- and post-tests of pronunciation, and how their validity and reliability was addressed is not stated. In addition, the fact that the tests were identical may have created a repetition effect. Another issue is that it appears that the teacher administered the weekly surveys, meaning that responses may have been overly positive as students sought to please their enthusiastic instructor. Finally, the lack of a control group makes it difficult to be certain that pronunciation improvement was due to the shadowing and tracking activities, or just having more exposure to natural French pronunciation via the listening materials provided. It should be noted that Willardson (2014) identifies the lack of control group as a key limitation of the study.

40)

Yavari, F., & Shafiee, S. (2018). Effects of shadowing and tracking on intermediate EFL learners' oral fluency. *International Journal of Instruction, 12*(1), 869-884.

Yavari and Shafiee (2018) investigated the effect of shadowing and tracking interventions on the fluency of intermediate level Iranian EFL students. They used purposive sampling to select 60 students out of a group of 112 intermediate level students, based on their proficiency level (participants had to have a Preliminary English Test score within one standard deviation of the mean). The 60 participants were all females aged 15-20, with Persian as an L1.

The researchers divided the 60 participants into four groups, each of which was to receive a different intervention: a shadowing only group (SG), a tracking only group (TG), a shadowing

and tracking group (STG) and a control group. Materials for shadowing and tracking were selected from videos contained in the intermediate textbook *Speak Now 1*. Participants received two 15-minute shadowing or tracking sessions per week, as part of normal classes, over a 5-week period.

To measure improvement in fluency, the researchers administered an identical pre- and post-test, which consisted of a semi-structured interview in which participants spoke about famous buildings in their town and holiday plans. Fluency was measured by two raters, according to syllables spoken per minute, and inter-rater reliability was high (.79 at pre-test, .81 at post-test).

To analyse whether shadowing, tracking, and shadowing and tracking each had an effect on fluency, the researchers conducted paired sample t-tests with results from the pre- and post-tests. SG, TG and STG all showed statistically significant improvements in pre- and post-test scores (p>0.05). To assess any differences between the experimental and control groups, one-way between groups ANOVA were used. There were no statistically significant differences between groups at pre-test level. There was, however, a statistically significant difference between the groups at post-test, with the three experimental groups showing significantly higher mean scores than the control group (p > 0.05). In addition, shadowing was significantly more effective than tracking (p = .000), and shadowing and tracking was significantly more effective than both shadowing alone and tracking alone (p = .000).

The results offer convincing support for the effectiveness of shadowing and tracking in improving fluency. However, there are several limitations of this study. Firstly, the authors do not specify the exact procedures used for either the experimental or control groups. The exact steps of shadowing and tracking are not specified, which is problematic as there are many different ways of carrying out these interventions. In addition, it is unclear if the control group watched the same videos as the experimental groups, or was not exposed in any way to the same content. This is important as the content of the videos seemed to be related to the content of the pre- and post-test and, as such, not being repeatedly exposed to the same vocabulary could have resulted in lower fluency scores. Other limitations are related to the rating of fluency. The fact that only one measure of fluency (syllables per minute) was used for the study is problematic, as fluency itself can be a controversial topic, with even native speakers showing slow rates of speech and hesitation. Additional measures like utterance length, composite measures, or correction based on L1 behaviour could have been used to complement the measurement. In addition, it could have been beneficial to include more raters in the study. Finally, using the same pre- and post- tests may have had a repetition effect on the fluency scores.

41)

Zając, M., & Rojczyk, A. (2014). Imitation of English vowel duration upon exposure to native and non-native speech. *Poznán Studies in Contemporary Linguistics, 50*(4), 495-514.

Zajaç and Rojczyk (2014) conducted a study to investigate whether the magnitude of imitation amongst L2 speakers of English varies according to whether the model speaker is a native or non-native speaker, and according to whether explicit instructions are provided before imitation.

Participants were 40 Polish learners of English, of whom 31 were females and nine of whom were males. All were first-year university students enrolled in the Institute of English, at the University of Silesia. Participants' English proficiency was B2, and their median age was 19. None had received instruction about the durational variability of English vowels as a cue to the voicing status of following consonants.

The experiment involved imitation of 24 monosyllabic English words with the front vowels (/ɪ æ ɛ/) flanked by word-initial /b/, /m/, or /s/ and word final /t/ or /d/. The participants imitated recordings of these words from a native speaker of Standard Southern British English and a qualified male phonetician imitating a Polish accent. The words in the model recordings were recorded at a natural speaking tempo with falling intonation. Also, vowel duration was measured and the PSOLA technique was used to average any measured differences in duration (to guarantee equal normalised durations of vowels before voiced and voiceless stops in the Polish model).

20 participants took part in a session in which target-model words were presented without specific instructions to imitate: participants were simply instructed to wait until the recorded voice stopped producing the word then read the word from the screen. The other 20 participants took part in a session in which they were instructed to imitate the words as faithfully as they could. In both sessions, the order of the presentation of the stimuli words was counterbalanced, so that 10 students heard the native model first, and 10 heard the Polish model first. In all cases, the approximate interval between imitation and the onset of the next word was 1 second. It should be noted that, because of the lack of simultaneity in listening and reading, and the fact that only single words were imitated, this study does not fit with the definition of shadowing as used in this review. This is especially so as shadowing is not defined anywhere in the literature review. In the author's defence, "shadowing" is only referred to once in the abstract of the study, which indicated that participants read words after exposure to model talkers in "shadowing conditions".

To analyse results, 10 words were excluded due to frequent mispronunciation by participants, leaving 14 for analysis. Vowel duration was then analysed in a mixed 2*3*2 ANOVA design, with one between-subject variable with two levels (instruction/no instruction) and two within subject variables, one with three levels (word list/exposure to native English/exposure to Polish English) and one with two levels (vowel duration before voiceless consonant/vowel duration before a voiced consonant).

Results indicated that participants produced significantly longer vowels when imitating both the native speaker and the Polish speaker, compared to their baseline productions. In addition, explicit instructions did not have a significant effect on the magnitude of imitation. From these results, the authors suggest that there was convergence towards the native speaker, and divergence from the non-native speaker. They also suggest that participants may focus more on the characteristics of the experimental procedure than on what the researcher actually tells them to do.

Overall, the study design seems robust and reliable. However, because the "shadowing condition" mentioned in the abstract is, in fact, imitation of isolated words, the results of this study are not relevant to this systematic review. The study is therefore not discussed in the results section.

42)

Zhang, K., & Peng, G. (2017). The relationship between the perception and production of non-native tones. *Interspeech 2017*, 1799-1803.

Zhang and Peng (2017) conducted a study to explore Mandarin speakers' perception and production of Cantonese tones after a 2-week training program.

Participants were 35 Mandarin speakers, who had never learnt Cantonese, and 17 native Hong Kong Cantonese speakers. All participants were undergraduates or postgraduates, and none had visual, audio, or cognitive problems. None had received any linguistic, psychological, or musical training. It should be noted that no information is provided about participants' age or gender.

The experiment involved a 2-week training program in which participants learnt to perceive and produce Cantonese tones. The training involved listening to and repeating 36 Cantonese tonal syllables, which were recorded by a female native speaker whose recordings were selected for their clarity. It should be noted that, from the training described, the exercises do not fit with shadowing as it has been defined for this review. Tasks only involved imitation, and did not involve listening and speaking simultaneously to longer utterances.

Data on perception and production before and after training was collected via pre- and post-tests for each skill. The perception test involved identification of tones of 36 syllables, each of which was repeated five times. The production test involved producing the pronunciation of 36 characters, which were presented with their corresponding Jyutping and tone letter.

First, pitch height and slope produced by all participants (Mandarin and Cantonese speakers) were calculated. Then, a native norm for each category was obtained by averaging 29 Cantonese speakers' (an additional 12 Cantonese speakers made the recordings for the training, and their recordings were included here) productions. The distance between a Mandarin subject's production and the native norm was calculated, with the smaller the distance, the better the pronunciation. Two-way repeated measure ANOVAs were used on the perception and production results to explore the relationship between session, tone, and group. Correlations were calculated between results of the perception tests and the production tests.

In identification tasks, there was a significant main effect for session and tone. The accuracy of the post-test was significantly higher than that of the pre-test. The subjects' performance varied greatly across tone categories, and there was a significant interaction between tone and group (i.e., Mandarin and Cantonese speakers' productions of tone). Both groups confused acoustically similar tones, but Cantonese speakers discriminated between these tones much better than the Mandarin speakers. For the production results, a significant main effect was found for session and tone, with the distance between participants' productions and native productions significantly reduced at post-test. Again, there was variation in accuracy of different tones, with a tone that existed in Mandarin produced much more accurately than others. Correlation analyses showed that perceptual accuracy was negatively and significantly correlated with the production distance in both pre- and post-tests, indicating that Mandarin learners who perceived the Cantonese tones better could also achieve a more native-like tone production. However, no significant correlation was found between performance change in the perception task and production task.

From these results, the authors conclude that Mandarin learners who can perceive Cantonese tones accurately may also produce these tones in more native-like ways. However, they note that, according to the data, an improvement in one modality (perception or production) may not lead to an improvement in the other. Indeed, in general, most Mandarin learners' perception, but fewer learners' production, were improved after training. Finally, they highlight that learning of tones appears to have been affected by their native tonal system.

The data collection and analysis in the study appears rigorous and trustworthy. However, despite this, the relevance of the study for this systematic review is very low, given its use of imitation rather than shadowing and the fact that the impact of shadowing was not the main focus of the research. For this reason, the study is not discussed in the results section of this dissertation.

43)

Zhang, X., Miyaki, T., & Rekimoto, J. (2016). WithYou: an interactive shadowing coach with speech recognition. *UIST'16 Adjunct*, 61-63.

Zhang et al. (2016) conducted a user study to explore the effectiveness of an automatic shadowing application they developed, called WithYou, compared to other methods of shadowing. The application is capable of adjusting the narration playback based on users' speech progress in real time. That is, if a user is lagging far behind the model speech, or if there is frequent mispronunciation, the playback is automatically rewound to the last punctuation before the user's mistake. They also explored user perceptions of their application.

Participants in the user study were seven graduate "non-native" students. No information is provided about the students L1, age, gender, or proficiency level. The students L2 is also not explicitly stated, although in the screenshots of the application it appears to be English. Participants shadowed with three different methods (conventional shadowing, WithYou, and a manual version of WithYou in which they could rewind the playback themselves), in a randomised order. No information is given as to what participants shadowed or how long the shadowing training lasted (i.e. whether the study was cross-sectional or longitudinal).

Unspecified measurements of performance before and after shadowing indicted that users improved more with the automatic WithYou system compared to other methods. In addition, the authors report that users with a relatively low shadowing performance before practising improved more than their counterparts with a higher level.

An unspecified subjective evaluation of the systems' usability indicated that shadowing with WithYou was considered easier than other methods and highly preferred to the conventional method.

From these results, the authors conclude that WithYou is capable of helping users, particularly those with relatively low shadowing performance, to improve learning efficiency by making shadowing easier to do. However, the data provided is not detailed enough to support these conclusions, especially given the small sample size of the user study. The lack of information about participants, data collection instruments, and analyses makes it very difficult to trust the findings provided. For this reason, although the study is highly relevant, its findings cannot make much of a contribution to this review.

44)

Zhang, X., Miyaki, T., & Rekimoto, J. (2020). WithYou: automated adaptive speech tutoring with context-dependent speech recognition. *CHI'20*. https://doi.org/10.1145/3313831.3376322

Zhang et al. (2020) developed a shadowing application called WithYou, which is able to evaluate learners' shadowing progress during use of the application. If performance in unsatisfactory, that is, if learners fall behind the model speech or mispronounce many words in a sentence, the application automatically lowers the difficulty of the shadowing text by adding pauses into the speech template. In their paper, they describe a user study in which they investigated how standard shadowing with a sound player compares to WithYou. The user study focused on how learners shadow with WithYou, WithYou's usability and cognitive load, and the learning output of WithYou compared to standard shadowing with a sound player.

Participants in the study were eight non-native speakers of English, with six females and two males in the experiment. Participants were aged 22-30 and had an intermediate level of English proficiency (TOEFL scores between 70-90 out of 120). It should be noted that participants' L1, and the country in which the study was carried out, are not mentioned.

Participants shadowed once via conventional shadowing and the WithYou application. Shadowing was well defined in the literature review and tasks seemed to focus both on immediate repetition and simultaneous speaking and listening. The texts that they shadowed in each method were of the same difficulty (grade 9-10 Dale-Chall Readability Test). Both the order of method of shadowing and the order of texts was randomised, and all possible combinations of order of method and text were included in the study. The measures the authors took to ensure that text difficulty and order did not influence the study are valuable.

Participants received training in both methods before they began. They then used the first method to shadow the text once, whilst listening to the speech. This shadowing was recorded as a pre-training level. Participants then shadowed three times, before completing a post-training recording. They also filled out a usability questionnaire, designed to measure both usability, via Likert 7-point scale questions, and the cognitive load of each method, via a NASA-TLX test. After the questionnaire, participants rested for two minutes and then repeated the procedure with the second method. After completing the procedure with the second method, participants took part in simple interviews to give their feedback on both ways of shadowing.

Repeat accuracy, the rate of words a learner correctly repeats in a sentence, was used to measure performance changes. When using WithYou, all eight participants improved their repeat accuracy in the post-training test. For the sound player, six participants improved their repeat accuracy, whilst two performed worse. Comparing the two methods, five participants improved more with WithYou than the sound player, two improved more with the sound player, and one showed no significant difference between methods. On average, WithYou helped participants improve approximately 14%, whilst standard shadowing helped participants improved approximately 2.7%. WithYou also appeared to help learners with low performance in particular, with the difference in pre- to post-test for three low performing learners ranging from 17.9%-25.6%. Standard shadowing did not appear to have this effect, with one low performer actually performing worse in the post test by making 12.8% more mistakes.

In terms of questionnaire results, all participants agreed that WithYou is easy to understand (n = 8) and to use for shadowing practice (n = 8). For the sound player, five participants agreed that it was easy to understand (n = 5), but most participants found it hard to do shadowing with (n = 6). All participants found the playback control in WithYou helpful (n = 8) and almost all agreed that it responded in time (n = 7). Most participants agreed that WithYou's evaluations met their own evaluations of their performance (n = 6) and all agreed that it made shadowing easier (n = 8).

In terms of interview results, six participants mentioned they could clearly see improvements using WithYou for three rounds, but not using the sound player. Six participants also mentioned that WithYou helped them stay motivated because they had little pressure to perform perfectly, as the system automatically re-starts the practice sessions. All participants preferred WithYou to the sound system, and all stated that the additional pauses made listening easier. However, some stated they would improve the system by adding a performance report when passing a session (n = 4) and that it might be better to differentiate the pauses introduced by the system from the pauses in the speech itself (n = 2). In addition, some (n = 2) indicated that they would like additional training sessions on the words they failed.

Results from the cognitive load evaluation indicated that participants had lower stress levels and higher levels of satisfaction with performance with WithYou compared to the sound player.

From these results, the authors conclude that WithYou is easier to use and has a lower cognitive load than shadowing with a sound player. In addition, they highlight that participants can achieve higher repeat accuracy improvement in WithYou than the sound player. These results should be taken with some caution for several reasons. Firstly, no calculations of statistical significance are given for the differences shown in repeat rates, meaning it is difficult to gauge the true extent of the differences between the two methods. Secondly, no information is given as to how interviews were conducted or analysed, making it difficult to gauge the trustworthiness of the information given. Finally, the sample size is very small, meaning that the generalisability of results is limited.