NOVA

University of Newcastle Research Online

nova.newcastle.edu.au

Matthews, Joshua; O'Toole, John Mitchell; Chen, Shen "The impact of word recognition from speech (WRS) proficiency level on interaction, task success and word learning: design implications for CALL to develop L2 WRS". Published in *Computer Assisted Language Learning* Vol. 30, Issue 1-2, p. 22-43 (2017)

Available from: <http://dx.doi.org/10.1080/09588221.2015.1129348>

This is an Accepted Manuscript of an article published by Taylor & Francis in *Computer Assisted Language Learning* on 20/01/2016, available online:
<http://www.tandfonline.com/10.1080/09588221.2015.1129348>

Accessed from: <http://hdl.handle.net/1959.13/1347287>

The impact of proficiency level on interaction, task success and word learning: design implications for CALL to develop L2 Word Recognition from Speech (WRS).

This paper reports on task interaction, task success and word learning among L2 learners of different levels of proficiency who used a CALL application previously shown to be effective in the development of L2 word recognition from speech (L2 WRS). Participants ($N = 65$) were categorised into three levels of proficiency according to their baseline ability to accurately recognise the phonological form of words in connected speech. All participants undertook the same web-based intervention task which involved reconstructing spoken text into the written form in a number of stages. In terms of task interaction, lower proficiency participants chose to listen to stimulus material more often and changed their efforts at reconstruction significantly more frequently than participants of higher proficiency levels. In relation to task success, lower proficiency learners were significantly less accurate overall and less immediate in their arrival at peak reconstruction accuracy than higher proficiency participants. A comparison of gain scores also showed that participants of different L2 WRS proficiency experienced significantly different word learning outcomes after using the application, with moderate proficiency participants significantly outperforming lower proficiency level learners in relation to academic word learning. A synthesis of the results is used to recommend design features of CALL for WRS development which will more adequately cater to learners of different proficiency levels. The paper concludes with some avenues for future research and some implications for the integration of computer-mediated L2 WRS development as a component of L2 listening development.

Keywords: Word recognition from speech; L2 listening; CALL design; Proficiency level

Introduction

Listening comprehension is an essential component of L2 language proficiency, but one which is particularly challenging for learners (Goh, 2000; Graham, 2006). The challenges of L2 listening can in large part be traced back to the unique features of spoken words. Of these features, arguable the most difficult for L2 learners is the transient nature of the speech signal. The transient nature of speech imparts the need for spoken words to be recognised quickly and at a rate which is not readily controlled by the listener (Hulstijn, 2007). The cumulative processing demands of extracting meaning from fluent speech can result in L2 listeners experiencing cognitive overload (Goh, 2000). Another feature of the speech signal which presents difficulty for the L2 learner is the blended nature of the words encoded in fluent utterances. Unlike written words on a page, demarcation between one word and the next is far less explicit, with fluent speech being characterised by the fusion of the end of one word with the beginning of the next. This feature results in clusters of blended words which have acoustic features significantly different to those of the individual words as they are pronounced in isolation (Field, 2003).

High proficiency listeners are readily able to deal with the transient and blended nature of spoken language (Field, 2008a, 2008b). For the high proficiency listener, the

recognition of words from speech occurs automatically and as such does not impose a significant burden on finite cognitive resources (Hulstijn, 2003). The automaticity of a high proficiency listener's recognition of spoken words is linked to a broad foundation of implicit phonological knowledge which has developed through extensive contextualised exposure to the target spoken language (Ellis, 2002). This store of implicit knowledge is important as it frees up cognitive resources for higher-level listening comprehension processes such as ascribing contextualised and nuanced meaning to the spoken message (Rost, 2002). Low proficiency L2 listeners are often unable to access and apply existing word knowledge while listening due to insufficient levels of automaticity in word recognition from speech (Goh, 2000). As a result, such L2 listeners must endeavour to apply strategies to extract meaning from a spoken message despite an inability to adequately perceive a sufficient proportion of its composite words.

The challenges experienced by L2 listeners provides motivation for L2 researchers and teachers to formulate, validate and refine approaches which alleviate these difficulties. A contemporary view of listening instruction is that it should comprise of two main branches: the development of listening strategies and the development of lower-level listening skills (Graham, Santos, & Vanderplank, 2010). Advocates of the value of strategy instruction contend that conscious monitoring of listening processes assists listeners to strategically accommodate for gaps in their linguistic knowledge and processing capacities (Graham, et al., 2010; Vandergrift, 2007; Vandergrift & Tafaghodtari, 2010). The second, and less broadly researched pedagogical branch, aims to develop lower-level listening skills. The most critical of these lower-level listening skills is word recognition from speech (WRS) (Rost, 2002), which is defined here as the ability to accurately map information encoded in speech onto the corresponding lexical units already stored in the mental lexicon. Advocates of the value of lower-level listening skill development assert that a minimum level of word recognition capability must be attained before higher-order listening strategies can be effectively applied (Graham, et al., 2010). Further, it is unclear to what degree the application of explicit knowledge contributes to the development of skills such as L2 WRS which are primarily dependent on automatic processing at the lexical level (Matthews, Cheng, & O'Toole, 2014). Notwithstanding the various debates which exist regarding the relative value of the two complementary approaches to listening instruction (Field, 2008b; Hasan, 2000; Osada, 2001), lower-level listening skills are of strong importance in successful listening comprehension (Author, in press; Tsui & Fullilove, 1998) and as such their systematic development is very important component of L2 listening pedagogy (Graham, et al., 2010).

The development of WRS is however particularly challenging in learning contexts within which immersion in the target language is unfeasible. This is the case as skills which depend on fluent application of word knowledge need to be developed through extensive engagement with contextualised spoken input (Ellis, 2002; Hulstijn, 2003). It is difficult to provide sufficient levels of contextualised target language input to L2 classroom learners, and a paucity of such input has negative effects on L2 word knowledge development (Jiang, 2000). CALL offers a means by which to alleviate this language learning difficulty by providing learners opportunities to engage with contextualised input in out-of-class contexts

through use of personal devices such as tablets, mobile phones and personal computers. Despite the effectiveness of CALL in the development of improved L2 WRS (Matthews et al., 2014), and its potential to alleviate the difficulties associated with the development of this skill in classroom learning contexts, the breadth of research focussing on this area is limited. This gap in existing knowledge extends to the body of empirical data which can be used to practically inform the improved design of CALL for L2 WRS. As one of the major advantages of CALL for L2 WRS is its potential to facilitate autonomous development of this skill outside of regular classroom time, it is important that the design features of such approaches adequately cater to an individual's learning needs.

The importance of L2 WRS in listening comprehension (Field, 2008b; Hulstijn, 2003; Rost, 2002), the practical difficulties in developing this skill and the current lack of empirical data which can be used to inform the design of CALL for the development of L2 WRS provide the primary motivations for the current study.

Literature review

The design of CALL for the development of word recognition from speech (WRS)

Theoretical frameworks for the development of lower-level listening skills in large part contend that improvements in L2 WRS depends on the provision of structured and extensive opportunities to engage with appropriate speech input (Field, 2003, 2008a, 2008b; Hulstijn, 2003; Wilson, 2003). In contrast to traditional dictation techniques which are teacher-centred, and provide an inadequate level of target speech input (Jarapur and Yamini, 1993; Kiany & Shiramiry, 2002; Rahimi, 2008), L2 WRS is most effectively developed through learner-centred text reconstruction tasks which offer multiple opportunities to listen to and reconstruct target text. Field (2008a, page 49) recommends that learners should be “given the opportunity to listen and re-listen as often as they wish until they are satisfied that they have achieved a correct segmentation”. Similarly, Hulstijn (2003) recommends that L2 WRS is most effectively developed through a learner-centred text reconstruction approach by which learners listen as often as they choose to short sections of speech as they reconstruct what is heard.

Broadly available digital technologies are well equipped to deliver spoken language in structured learning environments, thus making CALL well suited for the development of lower-level listening skills (Vandergrift, 2007). Past studies have provided empirical evidence for the general functionality of CALL in assisting learners of various languages, including Korean, Dutch and English, to make stronger connections between spoken and written word forms (Grezel & Sciarone, 1994; Matthews & O'Toole, 2013; Pyun & Lee-Smith, 2011). Early theoretical frameworks specifically aimed at computer-mediated development of L2 WRS (Hulstijn, 2003) have since been operationalised in real learning contexts with promising results. Matthews & O'Toole (2013) developed and investigated a web-based application with a design based on the second language acquisition (SLA) constructs of input, output and feedback. A single-group pre-test/treatment/post-test research design was applied in order to investigate the effectiveness of the text reconstruction

approach for the development of L2 WRS among a group of 33 learners of English as an L2. A feature of the application was the learner-centred control of the timing and number of opportunities to listen to the target text. A battery of word recognition tests indicated that statistically significant levels of improvement in the ability to recognise words from speech occurred among the participants. Of importance also was the finding that participants had a positive attitude towards the usefulness of the student-centred text reconstruction tasks.

Matthews, Cheng and O'Toole (2014) provide the clearest evidence for the effectiveness of learner-centred, text reconstruction tasks on the development of L2 WRS. A pre-test/treatment/post-test research design with a treatment and control group was implemented among a group of 96 tertiary level Chinese learners of English. The treatment group of 65 participants undertook the computer-mediated text reconstruction tasks, whereas a control group of 31 participants did not. Carefully validated tests which tapped L2 word recognition from speech indicated that, when compared to the control group, those in the treatment group had significantly greater mean L2 WRS gain scores. Results also indicated that the number of times participants elected to listen to the stimulus listening material had a measurable and positive effect on the magnitude of word learning outcomes experienced by a majority of the treatment group. This study provides very clear empirical support for the pedagogical value of tasks which provide learners repeated opportunities to listen to and attempt to reconstruct spoken language. However, a major criticism of this study is that it did not offer substantive discussion in regard to the need to differentiate design elements to cater to individual difference. The value of CALL for the development of L2 WRS is strongly linked to its utility in the delivery of out-of-class learning opportunities. Accordingly, it is important that such approaches have design features which respond adequately to the individual differences evident between learners. Gathering data drawn from learners involved in the use of a CALL application aimed at developing WRS, and using that data to inform the design of future iterations which cater to individual learner differences, is the central goal of the current paper

A focus on the learner in the design of CALL for listening development

Despite findings that design decisions have clear impacts on the success of CALL (Cárdenas-Claros, 2014), design research has been underrepresented in CALL literature in recent times (Felix, 2005; Hémar & Cushion, 2006). As the roll-out of digital resources which are of potential value for the development of L2 listening continues (Robin, 2007), so too does the need to more clearly understand how design decisions may impact on the effectiveness of these resources in real language learning contexts (Chapelle, 2009; Garrett, 1991). Using SLA theory to make CALL design decisions is one strand of the CALL design literature (Chapelle, 2004, 2009). However, in order to “prepare a strong and workable design, the designer must be sensitive to individual learner characteristics and the learning context” (Levy & Stockwell, 2006, page 36). Thus although SLA theory is a solid foundation from which the design process can begin, successful elaboration from such theoretical bases requires a careful consideration of how the user interacts with and benefits from CALL and how this information can be effectively applied in CALL design (Hémar & Cushion, 2001, 2006).

Various research paradigms have been used to inform user-centred design approaches in CALL for listening. These include the user and other stakeholders working collaboratively to establish design imperatives for CALL (participatory design) (Cárdenas-Claros, 2014), use of theoretical frameworks to guide the design of help options (Cárdenas-Claros & Gruba, 2013) and qualitative investigation of user's attitudes to various CALL design features (Cárdenas-Claros & Gruba, 2012). More closely aligned with the research approach adopted here are those studies which have used quantification of user interactions with established computer-mediated listening applications to draw conclusions about pedagogy and design (Chen, Zhang, & Liu, 2014; Grgurović & Hegelheimer, 2007; Hegelheimer & Tower, 2004; Roussell, 2011; Smidt & Hegelheimer, 2004). This approach to CALL design acknowledges that effective design should be iterative in nature and that revisions to existing designs should be guided by investigations of actual learner use of CALL systems (Heift, 2006).

Among the recurring themes in this research is the strong influence of individual difference on the way learners interact, use and benefit from CALL designed to improve L2 listening. Of the dimensions of individual difference investigated as part of research into CALL for listening, differences in learner proficiency is among the most widely addressed. Previous studies have related varying levels of learner proficiency to aspects of users' experience with CALL for listening including: the application of different listening strategies, use of help options and task success. A majority of these studies have concluded that learner proficiency has a significant impact on the way learners use and benefit from CALL for listening. For example, investigations have concluded that learner proficiency level had a significant impact on the types of strategies employed by listeners (Roussell, 2011; Smidt & Hegelheimer, 2004). Proficiency level has also been shown to impact the way learners access CALL help options, with higher proficiency learners using help options more effectively than lower proficiency learners. Additionally, proficiency level has also been found to play a role in the degree of success experienced by learners as they engaged with CALL for listening (Grgurović & Hegelheimer, 2007; Hegelheimer & Tower, 2004). This finding suggests that lower proficiency learners should be given increased levels of scaffolding in order to support higher levels of listening task success. However, specific recommendations on the nature of that scaffolding or recommendations on how that scaffolding may be effectively delivered to CALL users is not well developed in the relevant literature (Smidt & Hegelheimer, 2004)

The developments summarised above highlight the need to carefully consider the relationship between a learner's proficiency level and the design features of CALL aimed at improving L2 listening skills. Previous research in this area has primarily focussed on the development of listening comprehension in a global sense, and as with a vast majority of other L2 listening research, has not specifically addressed the recognition of words from speech as a central research objective (Broersma & Cutler, 2008). As such, it is unclear how previous research relates to the way learners of different proficiency levels will interact with CALL specifically designed to develop L2 WRS. Of the few CALL investigations which have focused on the development of L2 WRS as a central research objective, none to our knowledge, have focussed on the impact of learner proficiency on learner interaction, task success and word learning outcomes in an effort to inform CALL design. As such,

fundamental questions remain about how dimensions of individual difference, such as learner proficiency, should be catered to by the design features of CALL aiming to improve L2 WRS. It is the goal of this research to begin filling this gap in the existing CALL literature.

Research questions

Our initial goal here is to determine the degree to which differences in word recognition proficiency impact on the way learners interact with an application designed to improve L2 WRS. Second, we seek to determine the impact of proficiency level on the way learners handle the intervention tasks presented. By doing this we seek to determine if proficiency level is associated with different levels of task success achieved by learners. Lastly, we seek to determine if proficiency level has a significant impact on the word learning outcomes experienced by those who use the application. To achieve these research objectives the following research questions will be addressed:

- (1) What is the impact of learner proficiency level on the manner by which learners interact with a CALL application for the development of L2 WRS?
- (2) What is the impact of learner proficiency level on the varying levels of task success experienced by learners during interaction with a CALL application for the development of L2 WRS?
- (3) What is the impact of learner proficiency level on the word learning outcomes attained by learners after interaction with a CALL application for the development of L2 WRS?

The findings from the following research questions will be synthesised as recommendations for the design of CALL for the development of L2 WRS. The immediate goal of this research is to provide empirical data which will pave the way toward the design of CALL for WRS L2 development which more adequately caters to learners of different proficiency level.

Methodology

Participants

Participants for this study were 65 members of two first-year undergraduate classes of the same English language course studying at a large university in China. Students were assigned to these classes according to their English language proficiency level as determined by general proficiency tests which were delivered to all first year students at the beginning of the university year. All participants reported Mandarin as their first language and were aged between 17 and 20. The group was comprised of 33 males and 32 females. Their mean self-reported duration of English language study was 9.4 years ($SD = 2.4$, min.= 4, max. 18).

Overview of the web-application

The application used in this study has been described in previous research and has been shown to be effective in the development of L2 WRS (Author, date a; date b). For the present study the application was set to deliver 8 monologues. The monologues had a mean word length of 67.3 (min. = 59, max. = 75) and contained general information relating to countries of Asia. As the content of the monologues was factual it was possible, with only a few exceptions, to present the content in the present tense. This was used as a mechanism by which to control the degree of syntactic complexity present between monologues. Each monologues was divided into 9 monologue sections. The division of the monologue sections was achieved through use of digital audio editing software (WavePad). The digital audio file of each monologue was imported into the editing suite in order to show a visual representation of the waveform of the monologue. This enabled the identification of the boundaries between the intonation units present within the monologues. Intonation units are utterances of connected speech which mark the natural rhythms for presenting idea while speaking (Rost, 2002). Pauses at the beginning and the end of intonation units were thus considered a valid criteria for dividing the monologues into sections. Each monologue section had a mean word length of 7.3 (min. = 6.3, max. = 8.3) and a mean speech rate of 3.5 syllables per second (min. = 3.3, max. = 3.8). To ensure accuracy, syllable rates were determined manually for each section. Mean syllable rates reported here are those of the continuous spoken language within each section (intonation unit) and are therefore not affected by the pauses between intonation units.

Table 1: An overview of monologue word length, speech rate and lexical frequency

Monologue	Total words	Mean number of syllables per second	Proportion of words in the 0-2K frequency range (%)	Proportion of words on the Academic Word list (%)	Proportion of words beyond the 2K frequency level and not on the Academic Word list (%)
1	70	3.4	68.6	5.7	25.7
2	72	3.5	70.8	11.1	18.1
3	67	3.8	80.9	4.4	14.7
4	59	3.3	76.3	8.5	15.2
5	75	3.7	78.7	8.0	13.3
6	66	3.6	78.8	6.1	15.1
7	62	3.4	74.2	8.1	17.7
8	67	3.3	83.6	7.5	8.9
Average	67.3	3.5	76.5	7.4	16.1

A contextual digital image was also sourced for each monologue section. The 8 monologues (72 monologue sections) were delivered via a web-based application accessed via a web-browser and computer. Participants listened to the monologue sections through headphones from an individual computer. The material associated with each monologue was delivered in the following cycle:

- Presentation of the static contextual image on-screen as the entire monologue section is heard once. During this phase learners do not reconstruct text.

- Presentation of each of the monologue's sections in sequence. During this phase learners had the opportunity to repeatedly listen to the sections and reconstruct the sections into the written form using the keyboard and mouse. Participants were able to click on an on-screen panel such that they controlled when the monologue section began playing. After clicking the "play" panel the monologue section played through without pauses. Participants were able to modify their efforts at reconstruction after each opportunity to listen. This modification involved participants using the keyboard and mouse to self-correct the text they had written in response to listening to the monologue section multiple times. Participants had the opportunity to listen to each sample of spoken discourse between 1 and 10 times. Participants were free to choose when to move on to the next monologue section and could do so by checking an onscreen checkbox and electing to submit their final attempt by clicking an on-screen panel.
- After submitting their final effort at reconstruction for each monologue section participants received on-screen feedback. This feedback included, a) the number of words correctly reconstructed, b) the concurrent presentation of the participant's final effort at reconstruction and the correct target text with discrepancies between the two highlighted in red, and c) an opportunity to hear the monologue section while the on-screen feedback remained in view.

Assessment of L2 word recognition from speech and word learning outcomes

Test structure

Word recognition from speech was measured before and after intervention with two equivalent partial dictation tests specifically designed to measure L2 WRS (Matthews, et al. 2014). The pre-test was used to determine the participants' baseline L2 WRS proficiency before the intervention. The post-test was used to determine the participants' level of WRS capability after the participants undertook the intervention.

These tests involve participants listening to a spoken stimulus sentence in order to fill in a single missing word from a contextual written sentence. The stimulus sentence was heard just once with approximately 4 seconds delay between the end of one stimulus sentence and the beginning of the next. The speech rate of the spoken stimulus was approximately 2.1 words per second with each stimulus sentence being on average between seven and eight words in length. As was the case for the speech used for the intervention monologues the spoken stimulus for the tests was recorded to MP3 format with professional quality digital recording equipment by an Australian native speaker of English.

Each test measured knowledge of 60 target words. The first 32 target words were high frequency words, while the last 28 were from the Academic Word List (Coxhead, 2000; Nation, 2001). Word list based structuring of the pre and post-test instruments enabled a measure of WRS gain scores to be established for both high frequency words and academic words. High frequency words were deemed to be appropriate targets, as they make up a

majority of the spoken language encountered by listeners (Adolphs & Schmitt, 2003; Nation, 2006). Academic words were considered appropriate target words as they lay beyond the traditional level of high frequency words and are therefore likely to be more challenging for language learners to recognise. Academic words are also of pedagogical value for tertiary level L2 learners as these words appear relatively frequently in academic texts. Word learning outcomes were quantified with gain scores (post-test minus pre-test) for total words and for two subsections of the word recognition tests: high frequency words and academic words.

Test validation

Following Matthews et al. (2014) a two-stage validation process was undertaken to ensure that the primary construct being tapped by the tests was the ability to recognise the phonological form of words as they are heard in connected speech. The first stage of this validation process involved asking a group of three native speakers to attempt to identify the target words from the blanks in the written contextual sentences without listening to the spoken stimulus. This process confirmed that the target words could not be systematically identified by referring to the contextual sentences within which the target words were embedded. The second stage of validation involved a new group of three native speakers undertaking the tests, this time while simultaneously listening to the spoken stimulus. All target words were correctly identified by the native speakers. This validation process was important in order to ensure that the tests were tapping test takers' ability to recognise word from speech through activation of bottom-up listening processes. Further, it ensured that the WRS proficiency levels required for high levels of test achievement were within the range of capabilities typical of very high proficiency target language users.

Test scoring

As the primary construct of interest is the ability to recognise words from speech, it was important that test takers were not penalised for minor spelling errors (Buck, 2001). In regards to the methodology of marking word recognition tests, the key concern is to limit the potential threat to construct validity caused by the requirement to represent words in the written form. In order to achieve this goal, a structured scoring rubric was used to ensure categorisation of minor spelling errors and assignment of marks for varying levels of word recognition was systematic (see Appendix 1). The rubric assigns full credit to responses which are written in the correct orthographic form. The rubric also assigns full credit to responses which include minor spelling errors which in no way impede the scorer's ability to recognise the target word. Half marks were assigned for words which could be readily recognised by the scorer, despite a degree of ambiguity introduced due to errors in the representation of the target word. The scoring rubric was piloted by two scorers and was shown to facilitate strong levels of inter-rater reliability. ($\kappa = .91, p < .01$). Scoring of the tests was undertaken by a research assistant who had been trained in the use of the piloted scoring rubric.

Test reliability

Analysis indicated that the pre-test ($\alpha=.89$) and post-test ($\alpha=.93$) both had good levels of internal consistency. Reliability coefficients between pre- and post-tests were determined through correlational analysis and were strongly and positively correlated; $r = .88$, $N = 65$, $p < .01$.

Procedures

After written informed consent to participate in the study was obtained from all 65 participants, each undertook the pre-test. Testing took place in two sessions on the same day. After pre-testing each participant received log-on credentials which allowed access to the web-based application. To ensure participants understood how to use the application, an in-class tutorial of its use was provided before intervention sessions began.

Intervention sessions took place in a scheduled weekly one hour listening laboratory class. All participants listened to all 72 monologue sections within the five 60 minute sessions scheduled for the completion of the listening tasks. After completing the intervention sessions each participant undertook the post-test. After completing the post-test, participants were sent an electronic summary of their task performance. These summaries provided a complete record of each written entry participants made in response to hearing the monologue sections. The summaries also showed how many times each participant listened to each of the monologue sections as well as the target text for each of the monologue sections.

Quantifying task interaction and task success

Task interaction

User task interaction data was automatically captured in the data base associated with the web-application and was therefore unobtrusively collected from the participants as they used the application. Two task interaction variables were determined for each of the participants: repetitions and modifications.

The first interaction variable, *repetitions*, was the number of times learners elected to listen to each monologue section. This variable was determined by calculating the average number of times each participant listened to the monologue sections, namely by determining the total number of repetitions for all the monologue sections divided by 72.

The second interaction variable, *modifications*, was defined as the mean number of times participants altered their efforts at text reconstruction for each monologue section. Each effort at reconstruction for a given monologue section, which was in any way different from the previous attempt at reconstructing the same monologue section, was considered one modification. Thus modification is a measure of the number of times participants self-corrected their effort to reconstruct the text. The total number of modifications was divided by 72 in order to determine the average number of modifications undertaken for each monologue section by each participant.

Task success

Two measures of task success were established for each of the participants: reconstruction accuracy and average peak reconstruction attempt.

Reconstruction accuracy is a measure of the total percentage of words correctly transcribed by participants as they attempted to reconstruct the spoken monologue sections into the written form. The total number of words correctly transcribed over the course of the 72 monologue sections was established and this value was then divided by the total number of words in the monologue sections and expressed as a percentage.

The second variable, *average peak reconstruction attempt*, was a measure of the average repetition number within which the participants attained the maximum degree of reconstruction accuracy for each monologue section. This variable provides information on how immediately each participant was able to attain their greatest number of correctly reconstructed words for each monologue section.

Analysis

The primary method of analysis was to compare the variables of task interaction, task success and word learning outcomes of the 65 participants who were categorised into three groups based on their initial WRS proficiency levels. This was achieved by using total WRS pre-test scores to establish three groups of equal size with relatively low, moderate and high WRS proficiency level. WRS proficiency level, as defined by these three categories, was the primary grouping variable used for analysis in this study. Analysis of variance (ANOVA) and post hoc testing were the main statistical analyses used to compare these mean values.

Results

The 65 participants were categorised into either relatively low, moderate or high WRS proficiency groups. The mean word recognition scores for the pre-test are shown in Table 2. Low proficiency participants ($n = 22$) were those who achieved 40% or less for total WRS, moderate ability participants ($n = 22$) were those participants who scored between 40% and 55.83%, and the high group ($n = 21$) were those who scored above 55.83%.

Table 2: Pre-test mean L2 WRS scores for the total, high frequency and academic words.

Test component	N	Maximum possible score	Mean raw score	SD	Mean WRS score (%)	SD
Total WRS	65	60	28.68	9.91	47.79	16.52
High frequency WRS	65	32	18.14	5.74	56.68	17.95
Academic WRS	65	28	10.53	4.63	37.64	16.54

To establish a picture of the low, moderate and high participants' initial level of L2 WRS proficiency, scores for total and test sub-sections, are shown in Table 3. ANOVA and post hoc tests were undertaken to determine if the differences between the three groups were statistically significant. Mean scores for each proficiency level, within each test component,

were significantly different ($p < .01$). The expectation that academic words would be more challenging than high frequency words was confirmed with each proficiency level achieving higher WRS scores for the high frequency words than for their corresponding scores for academic words.

Table 3: Mean scores for total, high frequency and academic words for participants from low, moderate and high L2 WRS proficiency levels

Test component	WRS proficiency group	N	Maximum possible raw score	Mean raw score	SD	Mean WRS score (%)	SD	Sig. (2-tailed)
	Low	22		17.52	5.35	29.20	8.91	
	Moderate	22		29.43	2.87	49.05	4.78	
	High	21		39.57	3.70	65.95	6.16	
	Low	22		11.84	3.86	37.00	12.06	
	Moderate	22		18.89	2.06	59.02	6.45	
	High	21		24.00	2.39	74.85	7.48	
	Low	22		5.68	2.12	20.29	7.57	
	Moderate	22		10.55	2.01	37.66	7.16	
	High	21		15.62	2.58	55.78	9.22	

** $p < .01$

Research question one

What is the impact of learner proficiency level on the manner by which learners interact with a CALL application for the development of L2 WRS?

Mean values of the interaction variables of repetitions and modifications for each proficiency category were compared with one-way ANOVA (Table 4).

Table 4: Comparison of two task interaction variables for three L2 WRS proficiency levels

Interaction variable	WRS proficiency group	n	Mean value of variable	SD	Sig. (2-tailed)	Effect size (η^2)
	Low	22	5.30	1.39		
	Moderate	22	4.20	.77		
	High	21	3.80	.91		
	Low	22	2.70	.63		
	Moderate	22	2.14	.56		
	High	21	1.85	.34		

** $p < .01$

One-way ANOVA showed a significant difference between the mean number of times participants of different WRS proficiency (low, moderate and high) elected to listen to each of the monologue sections (repetitions), $F(2,62) = 11.54$, $p < .01$, $\eta^2 = .27$. Post hoc comparison with Tukey's Honest Significant Different (HSD) tests indicated that differences between mean repetitions for the low WRS proficiency group ($M = 5.30$, $SD = 1.39$) and the moderate WRS proficiency group ($M = 4.20$, $SD = .77$) and the low WRS group and the high WRS proficiency group ($M = 3.80$, $SD = .91$) were significant ($p < .05$). Mean differences in repetitions between the moderate and high groups were not significant to the .05 level. The general trend observed in the data is an inverse relationship between WRS proficiency level

and the number of times participants elected to repeatedly listen to the aural stimulus. A Pearson product-moment correlation coefficient affirmed the negative relationship between WRS proficiency level and the number of times participants repeatedly listened, $r = -.56$, $N = 65$, $p < .01$. Further, the large effect size ($\eta^2 = .27$) indicates the high practical significance of WRS proficiency level on this interaction variable. This appraisal of the magnitude of effect size is based on the guidelines of 0.01 = small, 0.06 = medium, 0.14 = large (Cohen, 1988)

At first glance these results appear contradictory to those of previous research which has investigated the task interaction of L2 learners with CALL designed to improve L2 WRS. This previous research found that learners of relatively low proficiency listened more times (6.00 repetitions) to stimulus material than those of relatively high proficiency (7.22 repetitions) (Matthews & O'Toole, 2013). However, the entire cohort from the aforementioned study was only able to successfully reconstruct approximately 46% of the target text while engaged in the L2 WRS development intervention, and thus had an average baseline proficiency far lower than those in the low proficiency category of the current study. Therefore, preliminary evidence does suggest a general trend of an inverse relationship between proficiency and the number of times learners select to listen to speech samples while engaged in L2 WRS development tasks.

ANOVA also showed a significant difference between the mean number of times participants of different WRS proficiency groups revised their efforts at text reconstruction (modifications), $F(2,62) = 14.09$, $p < .01$, $\eta^2 = .31$. Again Tukey's HSD tests indicated that differences between mean modifications for the low WRS proficiency group ($M = 2.70$, $SD = .63$) and the moderate WRS proficiency group ($M = 2.14$, $SD = .56$) and the low WRS proficiency and the high WRS proficiency group ($M = 1.85$, $SD = .34$) were significant ($p < .05$). As with repetitions, mean differences in modifications between the moderate and high groups were not statistically significant. The general trend observed in the data was the lower proficiency participants revised their efforts at reconstruction more than those of higher proficiency level. A large effect size was evident ($\eta^2 = .31$), as too was a negative correlation between WRS proficiency level and the number of times participants modified their efforts at text reconstruction ($r = -.58$, $N = 65$, $p < .01$). Again this affirmed the inverse relationship between proficiency and the number of modifications undertaken and also highlighted the practical significance WRS proficiency had on this interaction variable.

In summary, these results suggest that listeners with moderate or high WRS proficiency listen to significantly fewer repetitions before moving on than do listeners of low proficiency. Further, listeners with moderate or high WRS proficiency make significantly fewer modifications to their effort at text reconstruction than do those of low proficiency. The number of repetitions and number of modification of those with moderate and high WRS proficiency were not significantly different.

Research question two

What is the impact of learner proficiency level on the varying levels of task success experienced by learners during interaction with a CALL application for the development of L2 WRS?

Mean values of the task success variables of reconstruction accuracy and peak reconstruction attempt were compared with one-way ANOVA (Table 5).

Table 5: Comparison of two variables of task success for three L2 WRS proficiency levels

Task success variable	WRS proficiency group	<i>n</i>	Mean value of variable	<i>SD</i>	Sig. (2-tailed)	Effect size (η^2)
	Low	22	68.46 %	11.63	.00**	
	Moderate	22	81.80 %	8.97		
	High	21	86.72 %	5.52		
	Low	22	3.86	.80	.00**	
	Moderate	22	3.29	.51		
	High	21	2.94	.43		

** $p < .01$

A one-way ANOVA showed a significant difference between the mean reconstruction accuracy of participants of different WRS proficiency groups, $F(2,62) = 23.35$, $p < .01$, $\eta^2 = .43$. Post hoc comparison with Tukey's HSD again indicated that differences in reconstruction accuracy for the low WRS proficiency group ($M = 68.46\%$, $SD = 11.63$) and the moderate group ($M = 81.80\%$, $SD = 8.97$) and the low WRS group and the high WRS group ($M = 86.72\%$, $SD = 5.52$) were significant ($p < .01$). Mean differences in reconstruction accuracy between the moderate and high groups were not significant to the .05 level. The general trend observed in the data was that WRS proficiency level was positively correlated with reconstruction accuracy, $r = .73$, $N = 65$, $p < .01$. As was expected, the variable of WRS proficiency had a very strong practical impact on reconstruction accuracy ($\eta^2 = .43$).

A one-way ANOVA also showed a significant impact of relative WRS proficiency grouping on peak reconstruction attempt, $F(2,62) = 12.58$, $p < .01$, $\eta^2 = .29$. As with the three previous analyses, Tukey's HSD indicated that differences in peak reconstruction attempt for the low WRS proficiency group ($M = 3.86$, $SD = .80$) and the moderate WRS group ($M = 3.29$, $SD = .51$) and the low WRS group and the high WRS group ($M = 2.94$, $SD = .43$) were significant ($p < .01$). Again mean differences in peak reconstruction attempt between the moderate and high groups were not significant to the .05 level. A negative correlation between WRS proficiency level and peak reconstruction attempt was evident, $r = -.56$, $N = 65$, $p < .01$. WRS proficiency also had a strong practical impact on the mean peak reconstruction attempt ($\eta^2 = .29$).

In summary, these results suggest that listeners with high and moderate proficiency L2 WRS proficiency levels reconstructed text more accurately and more immediately than those of low proficiency level. However, the difference in reconstruction accuracy and peak reconstruction attempt between those of moderate and high WRS proficiency groups was not statistically significant.

Research question three

What is the impact of learner proficiency level on the word learning outcomes attained by learners after interaction with a CALL application for the development of L2 WRS?

A one-way ANOVA showed no significant difference between the word learning outcomes for total words or high frequency words attained by participants from the three WRS proficiency levels. In contrast, a statistically significant difference between the word learning outcomes for academic word gain scores was evident, $F(2,62) = 3.91, p < .05, \eta^2 = .11$.

Tukey's HSD indicated that mean academic word gain scores between the low ($M = 4.66, SD = 4.12$) and moderate proficiency level groups ($M = 7.45, SD = 3.10$) were significant ($p < .05$). Mean academic word gain scores between low and high and moderate and high groups were not statistically significant to the .05 level. The moderate to high effect size ($\eta^2 = 0.11$) suggests that WRS proficiency grouping had a practical impact on mean academic word gain scores.

Table 6: Comparison of total, high frequency and academic L2 WRS gain scores for three WRS proficiency levels

Test component	WRS proficiency group	<i>n</i>	Raw gain score (words)	<i>SD</i>	<i>Sig.</i> (2-tailed)	<i>Effect size (η^2)</i>
	Low	22	10.48	6.35		
	Moderate	22	13.14	5.39		
	High	21	10.10	3.40		
	Low	22	5.82	3.67		
	Moderate	22	5.68	3.85		
	High	21	4.41	2.67		
	Low	22	4.66	4.12		
	Moderate	22	7.45	3.10		
	High	21	5.95	2.49		

* $p < .05$

Discussions

Implications for the design of CALL for the development of L2 word recognition from speech

Although the participants of this study had been assigned to classes according to their general English language proficiency level, results from the partial dictation pre-test clearly showed that among this group, a significant variation in WRS proficiency existed. This finding is significant as it indicates that general categorisations of proficiency level based on a broader array of language macro-skills may not be sensitive to specific domains of language proficiency such as L2 WRS. As shown in the results of this study, these differences in word recognition proficiency have significant impacts on the manner by which learners interact with processing tasks, as well as the degree to which learners are able to handle those tasks successfully. Further, the differences in WRS proficiency had a measurable effect on the specific word learning outcomes experienced by learners. This finding suggests the merits of a differentiated approach to L2 WRS development according to an individual's L2 WRS

proficiency. However, from a practical perspective it is difficult for an individual teacher to facilitate such differentiated learning experiences without the benefit of well-designed CALL.

It is important to note that a single partial dictation test robustly predicted differences in task interaction, task success and word learning outcomes. Although pre and post-testing was undertaken in pen and paper format for the purposes of this study, partial dictation tests are able to be effectively administered with computers (Coniam, 1996, 1998). Computerised partial dictation tests, similar in format to those used as part of the present study, represent a useful design feature of computer-mediated development of L2 WRS. The administration of such tests at the beginning of the intervention process could provide a quantification of WRS proficiency and be used to automatically align learners of differing proficiencies with appropriate listening texts. Additionally, administration of computerised post-testing after intervention phases may be a convenient way to obtain data relating to the WRS improvements experienced by learners.

A number of the results from the present study suggest the importance of aligning learners of different WRS proficiency levels with appropriate listening materials. For example, despite the opportunity to listen multiple times, low proficiency learners were on average only able to accurately reconstruct approximately 68 percent of the target text. This level of word recognition is well below the level necessary for the comprehension of spoken discourse (van Zeeland & Schmitt, 2013). In light of recommendations that spoken texts containing only a few unknown words are appropriate for the training of word recognition skills (Hulstijn, 2003), it would seem that for the low proficiency participants involved in this study, the listening texts used were too challenging. How might this information usefully inform the design of CALL which aims to cater to the individual learning requirements of those with a relatively low capability to recognise words from speech? Automated quantification of task success, such as reconstruction accuracy, can be used as a method of formative assessment. Such measures of task success can then be used as a point of comparison with pre-selected achievement benchmarks. Such comparisons can be used to trigger the automated alignment of the learner with an appropriate listening text for the training of L2 WRS. For the purposes of this study, in order to standardise the listening texts encountered by all participants, a single set of texts were used. However, it is suggested that a range of listening materials could be made available to learners at different levels of proficiency. Just as individual difference in reading proficiency is catered to by graded readers of differing degrees of complexity, so too could the needs of learners of different WRS proficiency be catered to by graded listening materials (Wilson, 2003). Although a labour intensive and technically challenging task, a number of levels of listening materials which aim to cater to learners of different WRS proficiency could be developed. A collection of listening materials with a known range of word frequency profiles, speech rates, and levels of semantic and syntactic complexity would be of strong value in this regard (Révész & Brunfaut, 2013). CALL users that are consistently unable to accurately recognise spoken texts with a known level of complexity, could be automatically directed towards less challenging texts. Empirical data relating to task interaction and task success presented here offers a starting point from which to loosely base performance standards. Although

speculative due to the preliminary nature of these findings, perhaps for those with linguistic attributes similar to those of the participants involved in the present study, a cut-off of an 80 percent accuracy level is a useful rule of thumb below which learners are provided listening texts with higher proportions of high frequency vocabulary, slower speech rates and reduced semantic and syntactic complexity.

The principles outlined to ensure spoken texts are not overly challenging could also be used to ensure that listening materials are not pitched at an unsuitably low level for those with higher levels of proficiency. For learners involved in the study presented here, accuracy tending towards the 85 to 90 percent level, achieved within four repetitions, may be a suitable trigger for access to more challenging spoken texts. Again it is suggested that the difficulty levels of texts be increased by reducing the relative proportion of high frequency words, increasing syntactic and semantic complexity, as well as by increasing speech rates. The relatively straightforward mechanism suggested here of using on-task measures, such as reconstruction accuracy, as a trigger for alignment of learner proficiency with suitable listening texts is an example of what Heift (2007) refers to as a micro-adaptive approach to the individualisation of CALL. The advantage of this approach is that it has the potential to cater to the temporal nature of learner's abilities. Providing the performance benchmarks are set at appropriate levels, this form of automated delivery of listening materials provides a feasible mechanism by which to ensure listening materials are neither too challenging nor too simple for the learning needs of CALL users.

Other aspects of the results reported here have the potential to inform the design of CALL for the development of L2 WRS. For example, assuming the number of times participants elect to listen to a section of spoken text reflects a listener's proficiency and likelihood of task success, then such measures of interaction could also be used to trigger the provision of individualised scaffolding. Low levels of reconstruction accuracy coupled with high levels of repetition, perhaps beyond five or six repetitions, could trigger the automated provision of scaffolding such as glosses, cloze style structures for the target utterances or by providing the onsets of low frequency words (Conaim, 1998; Hartwell, 2010; Jones, 2006). In the instance where very high levels of scaffolding were required, an array of composite words and distractors could be presented in "drag and drop" panels to assist in the text reconstruction process. Such scaffolding would provide learners a highly structured form of help which would greatly reduce task difficulty, provide opportunities to effectively apply compensatory strategies and thus increase the likelihood of task success. Individualised provision of help options which are triggered by learners' relative success and task interaction, in our opinion hold great promise. If learners are explicitly aware, for the sake of argument, that needing more than 5 repetitions to reconstruct a text to an 80 percent accuracy level is an indication that they require "help", then it may be the case that learners will be more likely to take advantage of the help options made available. Such micro-adaptive provision of help may be particularly beneficial for lower proficiency learners who typically use help options ineffectively or not at all (Grgurović & Hegelheimer, 2007; Hegelheimer & Tower, 2004; Pujolà, 2002).

The finding that the academic word learning improvements of moderate proficiency participants were significantly greater than those of low proficiency participants also has implications for CALL design. The finding suggests that design mechanisms which effectively guide learners toward appropriate listening materials and adequately scaffolded learning tasks, may be important in enhancing word learning outcomes. Although speculative, the listening material and learning tasks used here appeared to be most suitable for those participants of moderate WRS proficiency. It seems likely that moderate proficiency participants had adequate WRS capabilities to meaningfully engage with the spoken input and at the same time also had sufficient scope for significantly greater word learning improvements.

Conclusions and future research

The measures of task interaction and task success investigated here are suggested as useful variables which can be used to trigger automated alignment of listening materials with learners of different WRS proficiency. Establishing benchmarks of performance for these measures of task interaction and success also holds promise in the automated delivery of performance based scaffolding. Ensuring learners of differing L2 WRS proficiency are aligned with the most appropriate listening materials and modes of support is an important component of future efforts to more effectively use CALL for L2 WRS development.

It must be emphasised that the guidelines put forward here are based on empirical data drawn from speakers of a common L1 background, Mandarin Chinese, and therefore may not be generalisable to speakers of other language backgrounds. A potential avenue for future research is to investigate the effect L1 background may have on L2 learners' interaction with listening materials of various known levels of complexity. Of further interest would be to ascertain the influence listening text complexity may have on the manner by which L2 learners interact with listening material used in the development of L2 WRS. Data drawn from such investigations would provide a more valid basis from which to recommend general guidelines for learners of a broader ranges of L1 backgrounds.

Achieving success in developing more refined approaches to L2 WRS development, especially those with effective triggers for micro-adaptive individualisation, has significance for listening pedagogy generally. Out-of-class computer-mediated learning which brings about effective and adequately individualised development of L2 WRS provides an attractive option for language educators. Adequate L2 WRS proficiency is essential for good listening comprehension, but its development in the classroom is challenging. As many aspects of language proficiency need to be formally addressed in the language classroom, there is typically insufficient time to facilitate adequate levels of WRS. Improved iterations of the application investigated here may not only have a significant role in enhancing learners' L2 WRS proficiency, but also play a role in reinventing the structure of listening pedagogy generally. Computer-mediated (or mobile device-mediated) WRS development could be the mainstay of independent out-of-class listening skill development. While in the classroom, time could be more fully dedicated towards other essential components of listening

pedagogy, such as the development of effective listening strategies and two-way interactive listening tasks.

It is suggested that CALL has the potential to deliver effective WRS development to learners in relatively autonomously, out-of-class contexts. However, crucial to the feasibility of this suggestion is the degree to which design features can cater to CALL users of different proficiency levels. This research has sought to provide empirical data which can be used to begin working towards this design objective. However, the design recommendations based on these data should be viewed as just one phase of an ongoing cyclical design process. Applied research investigating the design features recommended here is necessary to assess their validity in real language learning contexts. For example, investigating the way learners of different proficiency levels interact with and derive benefit from CALL for WRS development which operationalise the recommendations put forward here would be of strong interest. Further, tapping deeper into the individual user experience through walkthroughs or think-aloud protocols obtained from learners as they use such CALL approaches would also provide data which could inform future design revisions.

Developing methods to improve the efficiency of L2 WRS assessment is also an important avenue of future research. Although manual modes of scoring of L2 WRS do provide strong levels of inter-rater reliability (Matthews & O'Toole, 2013; Matthews, Cheng & O'Toole, 2014), a major limitation of the approach is that it requires ongoing reference to a detailed scoring rubric and as such is labour intensive and time consuming. Although it was not always a straightforward task to differentiate between an error in L2 WRS and an error in spelling, scoring rubrics such as those used as part of this research do enable such decisions to be made in a systematic and reliable manner. The development of algorithms which apply principles such as those reflected in the scoring rubric used in this research, may provide a useful starting point from which computers can be used to quantify L2 WRS in a valid, reliable and time effective manner.

The results presented in this paper suggest that variables such as repetition, modification and reconstruction accuracy provide useful indicators for different levels of listening proficiency. CALL applications allow the capacity to respond differentially to the levels so indicated. Such differential responses could include automatic guidance down pathways involving listening texts of differing difficulty and provision of varying degrees of task scaffolding beyond the aural modality. It may seem intuitive to suggest that learner interactions might scaffold CALL design, but these results provide empirical support for that intuition and such support could make appropriate developments more likely.

References

Author (date a)

Author (date a)

Author (in press)Adolphs, S., & Schmitt, N. (2003). Lexical coverage in spoken discourse. *Applied Linguistics*, 24, 425-438.

- Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, 36, 22-34.
- Buck, G. (2001). *Assessing listening Cambridge*: Cambridge University Press.
- Cárdenas-Claros, M. S. (2014). Design considerations of help options in computer-based L2 listening materials informed by participatory design. *Computer Assisted Language Learning*, 1-21. doi: 10.1080/09588221.2014.881385
- Cárdenas-Claros, M. S., & Gruba, P. (2013). Decoding the "CoDe": a framework for conceptualizing and designing help options in computer-based second language listening. *ReCALL*, 25, 250-271.
- Cárdenas-Claros, M. S., & Gruba, P. A. (2012). Listeners' interactions with help options in CALL. *Computer Assisted Language Learning*, 27, 228-245.
- Chapelle, C. (2004). Technology and second language learning: expanding methods and agendas. *System*, 32, 593-601.
- Chapelle, C. (2009). The relationship between second language acquisition theory and computer-assisted language learning *The Modern Language Journal*, 93, 741-753.
- Chen, L., Zhang, R., & Liu, C. (2014). Listening strategy use and influential factors in Web-based computer assisted language learning. *Journal of Computer Assisted Learning*, 30, 207-219.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. New Jersey: Lawrence Erlbaum.
- Coniam, D. (1996). Computerized dictation for assessing listening comprehension. *CALICO Journal*, 13, 73-85.
- Coniam, D. (1998). Interactive evaluation of listening comprehension: how the context may help. *Computer Assisted Language Learning*, 11, 35-53.
- Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, 34, 213-238.
- Ellis, N. (2002). Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143-188.
- Felix, U. (2005). Analysing Recent CALL Effectiveness Research—Towards a Common Agenda. *Computer Assisted Language Learning*, 18, 1-32.
- Field, J. (2003). Promoting perception - lexical segmentation in L2 listening. *ELT Journal*, 57, 325-334.
- Field, J. (2008a). Revising segmentation hypotheses in first & second language listening. *System*, 36, 35-51.

- Field, J. (2008b). *Listening in the language classroom*. Cambridge: Cambridge University Press
- Garrett, N. (1991). Technology in the service of language learning: trends and issues *Modern Language Journal* 75, 74-101.
- Goh, C. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, 28, 55-75.
- Grezel, J., & Sciarone, A. (1994). Computer testing of listening comprehension. *Computer Education*, 23, 125-132.
- Graham, S. (2006). Listening comprehension: the learners' perspective. *System*, 34, 165-182.
- Graham, S., Santos, D., & Vanderplank, R. (2010). Strategy clusters and sources of knowledge in French L2 listening comprehension. *Innovation in Language Learning and Teaching*, 4, 1-20.
- Grgurović, M., & Hegelheimer, V. (2007). Help options and multimedia listening: students' use of subtitles and the transcript. *Language Learning & Technology*, 11, 45-66.
- Hasan, A. (2000). Learner's perceptions of listening comprehension problems. *Language, Culture and Curriculum*, 13, 137-153.
- Hartwell, L. (2010). Impact of software design on on-line text reconstruction. *System*, 28, 370-378.
- Hegelheimer, V., & Tower, D. (2004). Using CALL in the classroom: Analyzing student interactions in an authentic classroom. *System*, 32, 185-205.
- Heift, T. (2006). Context-sensitive *Help* in CALL. *Computer Assisted Language Learning*, 19, 243-259.
- Heift, T. (2007). Learner Personas in CALL. *CALICO Journal*, 25, 1-10.
- Hémard, D., & Cushion, S. (2001). Evaluation of a Web-based language learning environment: the importance of a user-centred design approach for CALL. *ReCALL*, 13, 15-31.
- Hémard, D., & Cushion, S. (2006). Software design and development: Improving CALL design by establishing an informed dialogue between designers, developers and learners. *Computer Assisted Language Learning*, 19(2-3), 105-107.
- Hulstijn, J. (2003). Connectionist models of language processing and the training of listening skills with the aid of multimedia software. *Computer Assisted Language Learning*, 16, 413-425.
- Hulstijn, J. (2007). Psycholinguistic perspectives on second language acquisition. In J. Cummins & C. Davidson (Eds.), *The international handbook on English language teaching* (pp. 783-796). Norwell: Springer.

- Jafarpur, A., & Yamini, M. (1993). Does practice with dictation improve language skills? *System*, 21, 359-369.
- Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics*, 21, 47-77.
- Jones, L. (2006). Effects of collaboration and multimedia annotations on vocabulary learning and listening comprehension. *CALICO Journal*, 24, 33-58.
- Kiany, R., & Shiramiry, E. (2002). The effect of frequent dictation on the listening comprehension ability of elementary EFL learners. *TESL Canada Journal*, 20, 57-63.
- Levy, M., & Stockwell, G. (2006). *CALL Dimensions: Options and Issues in Computer-Assisted Language Learning*. Mahwah: Lawrence Erlbaum Associates, Inc.
- Matthews, J., Cheng, J., & O'Toole, J.M., (2014). Computer-mediated input, output and feedback in the development of L2 word recognition from speech. *ReCALL*, 1-19. doi:10.1017/S0958344014000421.
- Matthews, J., & O'Toole, J. M. (2013). Investigating an innovative computer application to improve L2 word recognition from speech. *Computer Assisted Language Learning*, 1-19. doi: 10.1080/09588221.2013.864315
- Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*: Cambridge University Press.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63, 59-82.
- Osada, N. (2001). What strategy do less proficient learners employ in listening comprehension? A reappraisal of bottom-up and top-down processing. *Journal of the Pan-Pacific Association of Applied Linguistics*, 5, 73-90.
- Pujolà, J. (2002). CALLing for help: researching language learning strategies using help facilities in a web-based multimedia program. *ReCALL*, 14, 235-262.
- Pyun, D., & Lee-Smith, A. (2011). Reducing Korean heritage language learners' orthographic errors: the contribution of online and in-class dictation and form-focused instruction. *Language, Culture and Curriculum*, 24, 141-158.
- Rahimi, M. (2008). Using dictation to improve language proficiency. *The Asian EFL Journal Quarterly*, 10, 33-47.
- Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition*, 35, 31-65.

- Robin, R. (2007). Commentary - learner based listening and technological authenticity. *Language Learning & Technology* 11, 109-115.
- Rost, M. (2002). *Teaching and researching listening*. Harlow: Longman.
- Roussell, S. (2011). A computer assisted method to track listening strategies in second language learning. *ReCALL*, 23, 98-116.
- Smidt, S., & Hegelheimer, V. (2004). Effects of online academic lectures on ESL listening comprehension, incidental vocabulary acquisition, and strategy use. *Computer Assisted Language Learning*, 17, 517-556.
- Tsui, A., & Fullilove, J. (1998). Bottom up or top down processing as a discriminator of L2 listening performance. *Applied Linguistics*, 19, 432-451.
- van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: the same or different from reading comprehension? *Applied Linguistics*, 34, 457-479.
- Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching*, 40, 191-210.
- Vandergrift, L., & Tafaghodtari, M. (2010). Teaching L2 learners how to listen does make a difference: an empirical study. *Language Teaching*, 40, 191-210.
- Wilson, M. (2003). Discovery listening – improving perceptual processing. *ELT Journal*, 57, 335-343.

Appendix 1. WRS Scoring rubric with rationale and examples

General Instructions

A rubric which outlines the principles of the scoring system is provided below. Examples which illustrate the guiding principles to be used for scoring word recognition from speech (WRS) are given. The example answers provided are not a comprehensive list. A list will need to be compiled by the marker and each unique answer will need to be considered on a case by case basis. As marking decisions are made, the list of previously marked words will need to be considered in order to work through problematic answers in a systematic manner. Knowledge of the spelling and sound system conventions of English will enable the marker to make a decision based on implicit knowledge of the target language words. At times reasoned decisions must be made based on the scoring rubric or previous marking decisions.

When scoring word recognition tests it should be remembered that:

1. The test is not a spelling test. However, it is important to ensure that there is compelling evidence that the phonological form of the word has been recognised by the test taker in order for a mark or half mark to be ascribed.
2. It will be necessary to make some decisions with some degree of subjectivity as part of the scoring procedure. It is therefore of primary importance to make such decisions in a systematic manner. Making systematic decisions based on the principles inherent to the rubric will assist in the reliability of the scoring process.

Score	Principle	Comments / Advise	Examples Target word → example answer
1.0	The word has been written in the correct orthographic form (both American and English spelling conventions accepted)	This is the easiest type of answer to score as there is no subjectivity involved.	Agriculture → agriculture Technology → technology Regional → regional attracts → attracts predicted → predicted relatively → relatively charmed → charmed recognise → recognize / recognise
1.0	Additional/incorrect consonant or vowel which doesn't affect phonological form of the word (marker must apply knowledge of the various ways sounds in English can be represented in various orthographic forms)	Subjectivity involved. First identify that there has been an orthographic error (spelling error). Then consider the orthographic error and decide whether, despite the error, the manner by which the word has been represented in writing enables the target word to be pronounced with the correct phonological form. As the recognition of the phonological form of the word is the target construct in this test, this type of orthographical error should not be penalised.	Agriculture → agraculture Technology → teknology Regional → reginal Attracts → attrakts Predicted → pridicted Relatively → relitively
0.5	Minor alteration to phonemic form of word through addition/deletion/substitution of vowels or consonants with some ambiguity of phonological form of the resultant word	Subjectivity involved. First identify that the manner by which the word has been represented in writing results in the target word being pronounced in an erroneous phonological form. However despite these minor errors, the marker is still satisfied that the test taker has recognised the target word. These words typically contain one incorrect vowel or consonant sound.	Agriculture → agriculter Technology → tacknollage / technilogy Diversity → daversity Immigrants → immgrance Significant → significan Temperature → tempature Mountains → montains
0.5	Incorrect conjugation / incorrect form of verb but clear evidence that the root word has been recognised	Some subjectivity involved. The core element of the target word has been correctly recognised however an error in inflection/word form has been made.	Located → locate / locating Occupied → occupy Predicted → predict Initiate → initiated Economy → economic Relatively → relative
0.5	Incorrect pluralisation / use of "s" / or failure to identify a plural	Simple to identify. If a word is correct except that it has been made a plural or a singular in error.	Contactors → contractor Processes → process Regard → regards
0.0	Significant alteration of vowels/consonants leading to significant phonemic alteration to word or significant ambiguity in the recognition of the target word – this may involve one of more significant vowel/consonant alterations	Subjectivity involved. The marker is not sufficiently satisfied that the test taker has clearly recognised the target word. This may involve a single or multiple incorrect vowel or consonant sounds which radically alters the phonological form of the word when pronounced.	Attracts → attarcts Occupied → occipied Ocean → occen Corporations → coprations Environment → inveroment Professional → perffessional Tourist → toris Villages → vellege Common → comond
0.0	No answer provided		