tesol QUARTERLY
tesol international association

Exploring Links Between Aural Lexical Knowledge and L2 Listening in Arabic and Japanese Speakers: A Close Replication of Cheng, Matthews, Lange and McLean (2022)

JOSHUA MATTHEWS
School of Education, University of New England
Armidale, Australia

AHMED MASRAI
Institute of Research and Consulting Services, Prince Sattam Bin Abdulaziz University
Al-Kharj, Saudi Arabia

KRISS LANGE
University of Shimane Matsue Campus
Matsue, Japan

STUART MCLEAN
Department of Business Administration, Momoyama Gakuin University
Izumi, Japan

EMAD A. ALGHAMDI
English Language Institute, King Abdulaziz University
Jeddah, Saudi Arabia

YOUNG AE KIM
Kyoto Seika University
Kyoto, Japan

YUKIE SHINHARA
Tokyo City University
Tokyo, Japan

SAORI TADA
Kwansei Gakuin University
Nishinomiya, Japan

Abstract

Aural lexical knowledge (ALK) is crucial for second language (L2) listening. Despite its importance, there is scant research that has validly explored the relationship between ALK and L2 listening across different English as a Foreign Language (EFL) contexts. In an effort to broaden this research base, the current study closely replicates a previous study, Cheng et al. (2022), which measured single-word ALK, phrasal verb ALK and L2 listening comprehension among participants with Chinese as a first language (L1). The current study administered the same instruments but did so among 147 Japanese and 131 Arabic-speaking English language learners. Results indicated that the capacity of ALK to predict variance in L2 listening for the Japanese group (R^2 = .38) was similar to that observed in the original study (R^2 = .42). However, the results for the Arabic-speaking group were very different to that of the original study and showed an unexpectedly strong relationship between ALK and L2 listening (R^2 = .92). Future research directions and pedagogical implications are discussed.

doi: 10.1002/tesq.3212

INTRODUCTION

The crucial role that L2 lexical knowledge plays in L2 listening relates to the unifying position that the lexicon holds at the interface between phonological and semantic components of a language system (Jackendoff, 2000). Hulstijn (2002) provides an explanation of the special importance of lexis in language learning:

> What gives the level of lexical units a special status is the fact that it is at this level that linguistic forms are matched with meanings. This makes them more amenable to conscious, metalinguistic reflection than formal units at the sublexical and supralexical levels. (p. 204)

The unity of form and meaning that occurs at the lexical level has important implications for L2 listening. Specifically, if a listener's bottom-up processes are sufficient to accurately recognize the phonological form of a target word, the associated semantic entries of that word can be activated in the mental lexicon. As words and their associated semantic representations are amenable to conscious reflection (Hulstijn, 2002), word recognition while listening facilitates purposeful application of top-down processes to build contextualized meaning (Field, 2008). Further, the activation of appropriate schema enables the listener to use an appropriate context to pre-emptively refine the lexical candidates likely to be heard in ongoing speech.

The current study focuses on the construct of *aural lexical knowledge* (ALK), which is defined as the capacity to map the phonological composition of words (*form*) that are encoded in speech onto an appropriate semantic representation (*meaning*). As the construct of ALK resides in the mental lexicon at the interface between knowledge of a word's phonological form and its semantic representation, this construct has strong relevance to the measurement of L2 listening. It is arguably the case that a listener's aural lexical knowledge modulates the capacity to successfully integrate top-down processes (i.e., largely non-linguistic knowledge, such as background knowledge) and bottom-up processes (i.e., largely implicit linguistic knowledge, such as recognition of sub-lexical linguistic units) (see, Figure 1).

Empirical evidence makes clear that lexical knowledge plays a crucial role in L2 listening processes. Although such evidence has been emerging for some time (Cheng & Matthews, 2018; Goh, 2000; Masrai, 2021; Matthews & Cheng, 2015), recent research that has applied structural equation modeling (SEM) is particularly important. Vafaee and Suzuki (2020), Wallace (2022) and Vandergrift and Baker (2015)

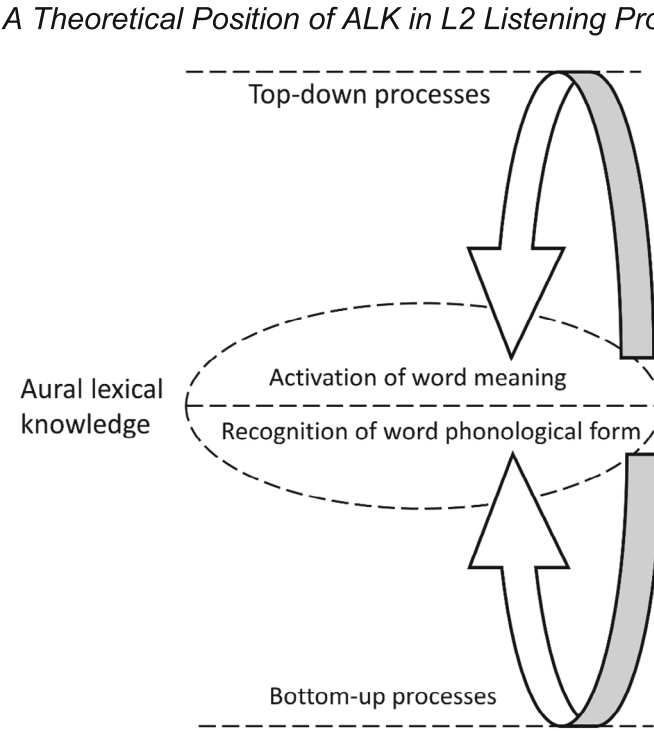

FIGURE 1. A theoretical position of ALK in L2 listening processes.

(reviewed below) all independently applied SEM to examine the relative impact of a range of factors hypothesized to be important to L2 listening. Across the three studies, a common finding was that L2 vocabulary knowledge was the variable most predictive of L2 listening comprehension. As Wallace (2022) reports, vocabulary knowledge was "... the strongest predictor of comprehension, suggesting that listeners with more vocabulary knowledge were better able to identify main ideas and details and make inferences based on information provided in speech than listeners with less vocabulary knowledge" (p. 29).

Despite its importance, there is scant research that has explored the relationship between ALK and L2 listening. Much of the previous research that has investigated links between L2 vocabulary knowledge and L2 listening has done so without validly measuring test takers' capacities to map phonological word form onto word meaning. A recent exception is Cheng, Matthews, Lange, and McLean (2022), which applied a meaning-recall test among Chinese learners of English as a foreign language (EFL). The current study closely replicates Cheng et al. (2022) in two different EFL contexts: EFL in Saudi Arabia and Japan. Cohorts of Arabic and Japanese-speaking EFL learners provide a valuable context within which to replicate Cheng et al. (2022) for a few reasons. Firstly, many learners of English possess Arabic or Japanese as their first language (L1). The total number of L1 Arabic speakers and Japanese speakers is approximately 360 and 125 million, respectively (WorldData, n.d.). Secondly, each language has distinct attributes (e.g., alphabetic versus logographic orthography) that may cast light on how L1 influence impacts upon the relationship between ALK and L2 listening. Further, due to attributes intrinsic to each language, Arabic and Japanese are at different language distances from the L1 language of the participants of the original study (Chinese). This will provide a useful point of comparison with the results of the original study and will enable the generalizability of the results reported in Cheng et al. (2022) to be interrogated across two distinct EFL contexts.

The Primacy of Aural Lexical Knowledge (ALK) in L2 Listening

Recent research drawn from EFL contexts has provided strong evidence for the primacy of ALK in supporting skilled L2 listening. Vafaee and Suzuki (2020) measured a range of variables among 263 EFL learners in Iran, including vocabulary breadth and depth, which were measured with a meaning-recognition format with aural stimulus. Of the variables tested, measures of vocabulary breadth and depth

were the most strongly correlated with the listening section of the International English Language Testing System (IELTS) (r = .78 & r = .80, respectively). Results from SEM indicated that both vocabulary knowledge and syntactic knowledge played a significant role in L2 listening success; however, vocabulary knowledge was a stronger predictor with an effect size approximately double that of syntactic knowledge. Vafaee and Suzuki (2020) affirm the importance of L2 vocabulary knowledge in L2 listening and acknowledge limitations of their vocabulary test by suggesting that "future studies should devise more valid ways for measuring aural [lexical] knowledge" (p. 405).

In a similar study undertaken among 226 Japanese EFL learners, Wallace (2022) investigated a range of factors including L2 vocabulary knowledge, topic knowledge, metacognitive awareness, working memory and attentional control. Listening was measured with the relevant section of the Test of English as a Foreign Language (TOEFL) and L2 vocabulary knowledge was measured with the Listening Vocabulary Levels Test (LVLT), which has a meaning-recognition format (McLean, Kramer, & Beglar, 2015). Listening comprehension test scores, which were categorized as those that measured understanding of detail, ability to inference answers and identify main ideas, showed moderate correlation with measures of L2 vocabulary knowledge (r = .42 to .57). The strength of correlation between L2 listening comprehension and L2 vocabulary knowledge was notably less than that determined by Vafaee and Suzuki (2020), which may suggest that the magnitude of this correlational relationship is influenced by factors such as test takers' L1. Wallace applied SEM and determined that "auditory vocabulary knowledge was most important for listening" (p. 36).

The findings of Vafaee and Suzuki (2020) and Wallace (2022) are supported by those of Vandergrift and Baker (2015), who investigated the role that different variables played in predicting L2 listening among 157 young learners of French. The variables investigated included L1 listening ability, L1 vocabulary knowledge, L2 vocabulary knowledge, auditory discrimination ability, metacognitive awareness and working memory capacity. The strength of correlation between vocabulary knowledge and L2 listening across three cohorts from three successive years was moderate (r = .51), but approximately double that of any other variable measured. Vandergrift and Baker (2015) conclude that "The robust role of L2 vocabulary in L2 listening comprehension [was] certainly the most significant finding" (p. 406). Vandergrift and Baker (2015) also reflect on their mode of L2 vocabulary measurement (i.e., the Peabody Picture Vocabulary Test) by noting that it may not tap lexical utility to the level of being able to recognize the target words in authentic contexts involving rapid and connected speech.

The studies reviewed above show that ALK is crucial to skilled L2 listening; however, the mode of measurement of ALK applied by Vafaee and Suzuki (2020) and Wallace (2022) has significant limitations. Each study used a test format modeled on the LVLT (McLean et al., 2015), which presents a target word in the aural modality in a non-defining sentence. The test taker must then recognize the spoken word and match it to the correct translated meaning written in the test taker's L1 (i.e., three distractors and one correct target, all presented in the L1). Because the target (and distractor items) are presented to the test taker, the LVLT is most accurately described as having a meaning-recognition format (see, example adapted from McLean et al. (2015), p. 743):

Aural stimulus: '*School. This is a big school.*' 1. a. 銀行 b. 海の動物 c. 学校 d. 家.

Tests with this item format provide evidence of the test taker's capacity to map the target word's phonological form onto a corresponding meaning (e.g., matching 学校 to *school*); however, there are important limitations. First, receptive vocabulary tests, by virtue of their format, are prone to guessing and influence from test strategies, which can result in the substantial overestimation of vocabulary knowledge (Kremmel & Schmitt, 2016; Stoeckel, McLean, & Nation, 2021). Additionally, the format of the test does not faithfully represent the way vocabulary knowledge is accessed during listening. Authentic listening demands the recognition of the phonological form of the target word in question and its association with an appropriate semantic representation held in the mental lexicon. Clearly this all needs to take place without the benefit of four alternative options presented in the written modality in the listener's L1. Thus, a major criticism of the meaning-recognition test format is that it does not tap a learner's capacity to independently access and employ lexical knowledge during fluent language use. Lexical employability in fluent language use, which is crucial for listening, is most validly measured at the meaning-recall level (Kremmel & Schmitt, 2016). A fundamental attribute of the meaning-recall test format is that test takers are "...asked to recall the word's meaning from memory without the assistance of a list of choices" (Stoeckel et al., 2021, p. 186).

The Original Study

Cheng et al. (2022) is the only study that we are aware of that has investigated the relationship between L2 vocabulary knowledge and L2 listening comprehension, through use of a meaning-recall L2 vocabulary test with aural stimulus. Cheng et al. administered two forms of L2 meaning-recall vocabulary test among 224 tertiary-level Chinese

EFL learners: one that measured knowledge of single words and one that measured knowledge of phrasal verbs. Both test types had 81 target items and entailed target word/s being articulated in isolation followed by the word/s being presented in a non-defining, fluently articulated contextual sentence. In order to evidence meaning recall, test takers had to write an appropriate L1 translation (e.g., Chinese) of the target item. Listening comprehension was measured with the listening section of the TOEIC. The strength of correlation between L2 listening and the two aural lexical measures was strong (single-word, $r = .65$) to moderate (phrasal verb, $r = .56$) (Plonsky & Oswald, 2014). Regression analysis indicated that single-word and phrasal verb ALK in combination could predict 42% of the variance observed in L2 listening comprehension. Of note was the relatively small but significant contribution of phrasal verb vocabulary knowledge to the predictive capacity of the regression model.

The Current Replication Study

The current study closely replicates the procedures used in Cheng et al. (2022). Porte and McManus (2019) assert that "one of the principle reasons for conducting a replication is to increase the confirmatory power of the original study" (p. 73). Porte and McManus go on to assert that when conducting a replication study care must be taken to systematically limit the variables that are modified. The current study seeks to replicate Cheng et al. (2022) by way of *close replication*. To do so, we will endeavor to keep the conditions of the original study as intact as possible and modify a single variable, namely the first language of participants. A close replication of this type is appropriate as a key objective of the current study is to interrogate the degree to which the results reported in Cheng et al. (2022) can be generalized to other EFL contexts. As is described in the relevant sections below, test instruments, test administration procedures, and test scoring protocols were identical to that of the original study. The target cohort of the original study was tertiary-level EFL learners in China, all of whom possessed Chinese as an L1. The current replication study targeted tertiary-level EFL learners possessing either Arabic or Japanese as an L1. The research questions of the current replication are fundamentally the same as those of the original except that these have been framed around a comparison of the different L1 language groups in question. The current study's design enables exploration of any differences that may be evident in the relationship between ALK and L2 listening comprehension between Arabic and Japanese L1 speakers. Research of this nature is important for the language teaching

community because speakers of different first languages are differentially sensitive to the structures of speech (Cutler & Otake, 2002) and these differences may implicate the need for different pedagogical interventions for different EFL cohorts.

Direct comparisons of Arabic and Japanese learners' spoken English word recognition difficulties in EFL contexts are not salient in the existing literature. However, Fender (2003) has compared Arabic and Japanese speakers' English written word recognition and contends that first language attributes do influence the processing skills noted among each language group. Arabic has an alphabetic orthography with a consistent grapheme–phoneme relationship (e.g., highly transparent), which is hypothesized to make Arabic speakers highly reliant on phonological processing skills (Bentin & Ibrahim, 1996). Conversely, Fender (2003) contends that the logographic writing system of Japanese (*kanji,* 漢字) encodes phonology far less transparently leading to Japanese L1 speakers developing "orthographic processing skills that are, in some ways, distinct and independent from phonological processing skills" (p. 294). With little existing research available, it is difficult to draw strong hypotheses about the impact that these noted cross-linguistic differences may have on the relationship between ALK and L2 listening among Arabic and Japanese L1 speakers. However, if L1 influence does have a systematic effect on the relationship between ALK and L2 listening, it is likely that this relationship among learners with Chinese and Japanese as an L1 will be more closely aligned than when compared to Arabic L1 learners. Both Chinese and Japanese have a logography with low grapheme–phoneme transparency and are arguably less linguistically distant from one another than when compared to the Arabic language.

Research Questions

1. For the Arabic and Japanese L1 groups, what is the strength of correlation between knowledge of aural single words and knowledge of aural phrasal verbs and L2 listening comprehension, and how do these relationships vary by L2 listening proficiency level?
2. For the Arabic and Japanese L1 groups, what is the relative knowledge level of aural single words and aural phrasal verbs, and how do these relationships vary by L2 listening proficiency level?
3. For the Arabic and Japanese L1 groups, to what degree do measures of aural single-word knowledge and knowledge of aural phrasal verbs predict variance observed in L2 listening comprehension scores?

METHOD

Participants

A total of 278 participants were involved in the study. All were tertiary-level learners of English as an additional language studying at four different educational institutions in Japan (n = 2) and Saudi Arabia (n = 2). Of the 147 Japanese-speaking participants, 94 were male and 53 were female. Of the 131 Arabic-speaking participants, 112 were male and 19 were female. Informed consent was obtained from all participants in accordance with the requirements of the research offices of the respective cooperating institutions. Scaled scores on TOEIC listening indicated that participants' L2 listening proficiency spanned A1 to C1 levels as described by the Common European Framework of Reference (CEFR). See Tables 1 and 2 for an overview and comparison. As is noted below, after data collection and analysis, it became clear that Arabic and Japanese-speaking participants had a different L2 listening proficiency distribution, which did reach a level of significance. For the purposes of comparing cross-linguistic influences on the relationship between ALK and L2 listening comprehension, this was not ideal. However, as multi-site transnational data collection is logistically challenging, the data set was considered valuable despite this limitation. As is explained below, steps were taken during analysis to unpack

TABLE 1
Breakdown of Arabic Speakers L2 Listening Proficiency

CEFR level	CEFR description	Number of participants	Mean TOEIC score (scaled)	*SD*
A1	Basic user	79	31.70	29.82
A2		21	177.14	44.06
B1	Independent user	13	360.77	35.64
B2		9	466.11	19.65
C1	Proficient user	9	495.00	0.00

TABLE 2
Breakdown of Japanese Speakers L2 Listening Proficiency

CEFR level	CEFR description	Number of participants	Mean TOEIC score (scaled)	*SD*
A1	Basic user	17	85.88	20.40
A2		92	189.24	40.23
B1	Independent user	31	318.23	33.50
B2		6	439.17	19.85
C1	Proficient user	1	495	0.00

the potential impact of differences in proficiency between the two language groups.

Listening Comprehension Test

As in Cheng et al. (2022), the listening section of the same version of the TOEIC (Educational Testing Service, 2018) was administered to measure English language listening proficiency. The TOEIC listening section requires approximately 50 minutes to complete and contains four parts in total containing 100 multiple-choice items. The listening section of the TOEIC has four parts: (1) choosing statements about photographs, (2) choosing correct responses to statements, (3) choosing correct information about conversations, and (4) choosing correct information about monologues. The same version of the TOEIC listening test was used in all of the data collection sites. Participants responded to the test stimulus with pen and paper and heard the aural stimulus on high-quality audio speakers. To interpret raw scores (which range from 0 to 100), with materials published by the test producers (Educational Testing Service, n.d.), scores were scaled from 5 to 495 using the same conversion matrix applied in Cheng et al. (2022) (See, Waikato Institute of Education, 2013).

Single-Word Meaning-Recall Vocabulary Tests

The single-word ALK test and audio files used in Cheng et al. (2022) were accessed via the Digital Repository of Instruments and Materials for Research in Second Language (IRIS) (https://www.iris-database.org/). The same list of target words and the same contextual sentences were used for each L1 cohort in the current study. The single-word ALK test consists of 81 items drawn from the spoken section of the Corpus of Contemporary American English (COCA). The frequency of occurrence of the single-word items decreases sequentially from one item to the next. The frequency range starts at 318.24 per million words (i.e., *won*) and ends at 1.00 per million words (i.e., *farmed*). For the meaning-recall format, the target word is first articulated in isolation and is then articulated as part of a fluently spoken utterance. The target sentences were recorded by a native English language speaker with a North American accent (see, Cheng et al., 2022). The test taker is required to write an acceptable L1 translation for the target word. Figures 2a,b exemplify the single-word meaning-recall item format for each L1 language group. The meaning-recall format measures the capacity of the test taker to recognize the phonological

(a)

Example of Single-word Meaning-recall Format (Arabic Response)

Aural stimulus: '*Bought: He bought them all'*
The Arabic L1 speaker writes: اشترى

(b)

Example of Single-word Meaning-recall Format (Japanese Response)

Aural stimulus: '*Bought: He bought them all'*
The Japanese L1 speaker writes: 買った

FIGURE 2. (a) Example of single-word meaning-recall format (Arabic response). (b) Example of single-word meaning-recall format (Japanese response).

form of the word and then map this onto an appropriate meaning. Kremmel and Schmitt (2016) contend that the meaning-recall test format offers the most robust measure of a test taker's capacity to independently access and employ lexical knowledge. A key advantage of the meaning-recall item format is that it negates concerns around the construct irrelevant variance implicit to meaning-recognition formats that provide L1 multiple-choice options, which may be correctly answered through strategic guessing (Stoeckel et al., 2021).

Phrasal Verb Meaning-Recall Vocabulary Tests

As with the single-word test, the phrasal verb ALK test used in Cheng et al. (2022) was accessed via the IRIS repository (https://www.iris-database.org/). The basic format of the phrasal verb test is the same as the single-word test. Each phrasal verb target item is matched closely to the frequency of occurrence of a corresponding single-word target to enhance each test's comparability to the other. For example, the phrasal verb frequency range starts at 316.37 per million words (i.e., *went on* matched with the single-word item *won*) and ends at 1.00 per million words (i.e., *set about* matched with the single-word item *farmed*). Figures 3a,b exemplify the phrasal verb meaning-recall item format for each L1 language group.

Scoring Procedures

Scoring procedures mirrored those described in Cheng et al. (2022). Two rater reference documents were developed for each rater group (see, Data sets S1 and S2). One rater group was composed of expert English language speakers with Arabic as an L1, and the other was composed of expert English language speakers with

(a)

Example of Phrasal Verb Meaning-recall Format (Arabic response)

Aural stimulus: '*Found out: They found out yesterday*'
The Arabic L1 speaker writes: اكتشف

(b)

Example of Phrasal Verb Meaning-recall Format (Japanese response)

Aural stimulus: '*Found out: They found out yesterday*'
The Japanese L1 speaker writes: 分かった

FIGURE 3. (a) Example of phrasal verb meaning-recall format (Arabic response). (b) Example of phrasal verb meaning-recall format (Japanese response).

Japanese as an L1. A bank of core and alternative acceptable translations for the single-word and phrasal verb target items were developed for each rater reference document. Members of the authorship team in each study location facilitated rater training and guidance around the scoring procedures. As with Cheng et al. (2022) raters were not constrained by the core and alternative suggested translations provided on the rater reference documents but were instructed to accept a range of translations that were appropriate for the target items according to the raters' experience and expertise. For each language, two raters rated all of the items. A primary rater's scores were used as the criterion measure and the other rater's scores were used to calculate interrater reliability, which was calculated with Cohen's Kappa (κ). Interrater reliability for the single-word and phrasal verb ALK tests for the Japanese-speaking cohort and the Arabic-speaking cohort were all very good ($\kappa > .80$).

Analysis

Initially, to determine if there was a difference in single-word and phrasal verb ALK between the Arabic and Japanese L1 speakers, the percentage of participants in each cohort that knew each item was first calculated. A Wilcoxon signed rank test was then applied on percentage distribution. To answer research question one, Pearson correlation was used to explore the relationship between ALK (single-word and phrasal verb) and L2 listening among each L1 language group. Further, to explore the magnitude of these correlational relationships for different proficiency levels, participants from each L1 group were categorized according to CEFR levels as Basic Users (A1 and A2) or Independent/Proficient Users (B1, B2 and C1) (Council of Europe, 2023).

To answer research question two, for each L1 group mean single-word and mean phrasal verb ALK scores were compared with *t*-tests

and effect sizes were estimated with Cohen's *d*. Again, the relative magnitudes of these mean scores were described for Basic and Independent/Proficient Users from each L1 group. Violin plots were used to visualize the differences in ALK mean scores. To answer research question three, hierarchical multiple regression analysis was used to determine the predictive capacity aural single-word knowledge and aural phrasal verb knowledge (predictor variables) had in explaining variance observed in L2 listening comprehension (outcome variable).

Visualizations were produced and analyses were undertaken using version 27 of the SPSS statistics software platform. For each multiple regression analysis undertaken, tolerance levels were all beyond .20, which indicates that multicollinearity was not problematic. Visual assessment of the Normal Q-Q plots, scatterplots and histograms of residual values indicated that the assumptions for analysis were adequately met for predictive modeling (see, Data set S3).

RESULTS

Research Question One - for the Arabic and Japanese L1 Groups, What is the Strength of Correlation Between Knowledge of Aural Single Words and Knowledge of Aural Phrasal Verbs and L2 Listening Comprehension, and How Do These Relationships Vary by L2 Listening Proficiency Level?

Preliminary analysis showed that there was a different distribution of baseline listening proficiencies between Arabic and Japanese groups. In relation to the CEFR framework, there was a higher proportion of A1 listeners in the Arabic group (see, Table 1) and more A2 level listeners in the Japanese group (see, Table 2). The difference between the two language groups was confirmed statistically. A Wilcoxon signed rank test revealed that single-word scores were significantly higher among the Japanese cohort (Md = 31.3, n = 147) compared to the Arabic cohort (Md = 21.9, n = 131), z = −4.09, p = .001, with a medium effect size (r = .46). Similarly, a Wilcoxon signed rank test also demonstrated that phrasal verb scores were significantly higher among the Japanese cohort (Md = 19.7, n = 147) compared to the Arabic cohort (Md = 9.3, n = 131), z = −6.27, p = .001, with a large effect size (r = .70).

Arabic Basic Users (A1/A2) had a mean TOEIC score of 62.25 (SD = 68.10) and Arabic Independent/Proficient Users (B1/B2/C1) had a mean TOEIC score of 430.32 (SD = 65.93). The Arabic cohort

in total had a mean TOEIC score of 149.35, with a very large level of individual variation (*SD* = 170.87).

Japanese Basic Users (A1/A2) had a mean TOEIC score of 173.12 (*SD* = 53.33) and Japanese Independent/Proficient Users (B1/B2/C1) had a mean TOEIC score of 341.97 (*SD* = 60.01). In total the Japanese cohort had a mean TOEIC score of 216.77 (*SD* = 92.30).

In response to research question one, the correlation between single-word ALK and L2 listening and phrasal verb ALK and L2 listening for the Arabic cohort was very large (see, Table 3) and for the Japanese group the correlation was moderate to large (see, Table 4). In terms of responding to the latter part of research question one, it can be seen that this level of correlation remained consistent across participants from each language group at each category of proficiency: Basic User and Proficient/Proficient User (see, Tables 5 and 6).

Research Question Two - for the Arabic and Japanese L1 Groups, What is the Relative Knowledge Level of Aural Single Words and Aural Phrasal Verbs, and How Do These Relationships Vary by L2 Listening Proficiency Level?

A summary of the mean scores for single-word and phrasal verb scores for the Arabic and Japanese groups are presented (see, Tables 7 and 8).

TABLE 3
Correlation Between the Three Test Variables for Arabic Speakers

Test	TOEIC listening	Single-word meaning recall	PV meaning recall
TOEIC listening	–	.936**	.955**
Single-word meaning recall		–	.944**
PV meaning recall			–

Note. *n* = 131, **Correlation is significant at the 0.01 level (2-tailed).

TABLE 4
Correlation Between the Three Test Variables for Japanese Speakers

Test	TOEIC listening	Single-word meaning recall	PV meaning recall
TOEIC listening	–	.599**	.522**
Single-word meaning recall		–	.723**
PV meaning recall			–

Note. *n* = 147, **Correlation is significant at the 0.01 level (2-tailed).

TABLE 5

Correlation Between the Three Test Variables for Basic (A) and Independent/Proficient Arabic Speakers (B/C)

Test	TOEIC listening	Single-word meaning recall	PV meaning recall
TOEIC listening	–	.708**	.825**
Single-word meaning recall	.601**	–	.768**
PV meaning recall	.713**	.689**	–

Note. [$n^{CEFR\ A}$ = 100 above diagonal (lighter shading), $n^{CEFR\ B/C}$ = 31 below diagonal (darker shading)]. **Correlation is significant at the 0.01 level (2-tailed).

TABLE 6

Correlation Between the Three Test Variables for Basic (A) and Independent/Proficient Japanese Speakers (B/C)

Test	TOEIC listening	Single-word meaning recall	PV meaning recall
TOEIC listening	–	.473**	.316**
Single-word meaning recall	.455**	–	.572**
PV meaning recall	.533**	.824**	–

Note. [$n^{CEFR\ A}$ = 109 above diagonal (lighter shading), $n^{CEFR\ B/C}$ = 38 below diagonal (darker shading)]. **Correlation is significant at the 0.01 level (2-tailed).

TABLE 7

Arabic Speakers: Single-Word and Phrasal Verb Scores for Total Cohort, Basic Users (A) and Independent/Proficient Users (B/C)

Group	*n*	Mean single-word score	*SD*	Mean phrasal verb score	*SD*
Arabic (total cohort)	131	23.93	20.44	18.84	16.08
Arabic (Basic Users)	100	13.86	8.78	10.96	6.82
Arabic (Independent/ Proficient Users)	31	56.42	11.28	44.26	9.65

TABLE 8

Japanese Speakers: Single-Word and Phrasal Verb Scores for Total Cohort, Basic Users (A) and Independent/Proficient Users (B/C)

Group	*n*	Mean single-word score	*SD*	Mean phrasal verb score	*SD*
Japanese (total cohort)	147	28.55	10.53	21.30	7.28
Japanese (Basic Users)	109	25.85	8.64	19.71	6.03
Japanese (Independent/ Proficient Users)	38	36.29	11.64	25.87	8.60

For the Arabic group, mean scores for single-word test scores were greater than those of the phrasal verb test scores. This was the case for the total cohort, the Basic Users, and the Independent/Proficient Users. It is noted that there was a very large standard deviation for the Arabic total cohort indicating large individual variation. The difference between the mean single-word and mean phrasal verbs scores for each category reached a level of statistical significance (Total cohort, $t(130) = 7.79$, $p < .001$, $d = 0.27$; Basic Users, $t(99) = 5.16$, $p < .001$, $d = 0.37$; Independent/Proficient Users, $t(30) = 8.07$, $p < .001$, $d = 1.16$) (see, Table 7).

The same trend was observed in the Japanese group, with mean single-word ALK scores being greater than the mean phrasal verb ALK scores. As with the Arabic group, this was the case for the Japanese total cohort, Basic Users and Independent/Proficient Users. This difference reached a level of statistical significance in all instances (Total cohort, $t(146) = 12.08$, $p < .001$, $d = 0.80$; Basic users, $t(108) = 8.95$, $p < .001$, $d = 0.82$; Independent/Proficient Users, $t(37) = 9.59$, $p < .001$, $d = 1.02$) (see, Table 8).

Of interest was a comparison of the mean single-word and mean phrasal verb ALK scores between the Arabic and Japanese participants. In order to account, to some degree, for the known differences between the L2 listening comprehension proficiency between the Arabic and Japanese speakers, these comparisons were undertaken within the Basic User and Independent/Proficient User categories.

Japanese Basic Users' mean single-word ALK scores ($M = 25.85$, $SD = 8.64$) were significantly greater than those of the Arabic Basic Users ($M = 13.86$, $SD = 8.78$); $t(207) = 9.95$, $p < .001$, $d = 1.38$. Similarly, Japanese Basic Users' mean phrasal verb ALK scores ($M = 19.71$, $SD = 6.03$) were also significantly greater than the Arabic Basic Users' mean phrasal verb ALK scores ($M = 10.96$, $SD = 6.82$); $t(207) = 9.84$, $p < .001$, $d = 1.45$.

However, Arabic Independent/Proficient Users' mean single-word ALK scores ($M = 56.42$, $SD = 11.28$) were significantly greater than those of the Japanese Independent/Proficient Users ($M = 36.29$, $SD = 11.70$); $t(67) = -7.23$, $p < .001$, $d = 1.75$. Similarly, Arabic Independent/Proficient Users' mean phrasal verb ALK scores ($M = 44.26$, $SD = 9.65$) were also significantly greater than the Japanese Independent/Proficient Users' phrasal verb ALK scores ($M = 25.87$, $SD = 8.60$); $t(67) = -8.27$, $p < .001$, $d = 2.01$.

Violin plots have been used to visualize the difference between single-word and phrasal verb ALK for Arabic and Japanese Basic Users (Figure 4) and for Arabic and Japanese Independent/Proficient Users (Figure 5).

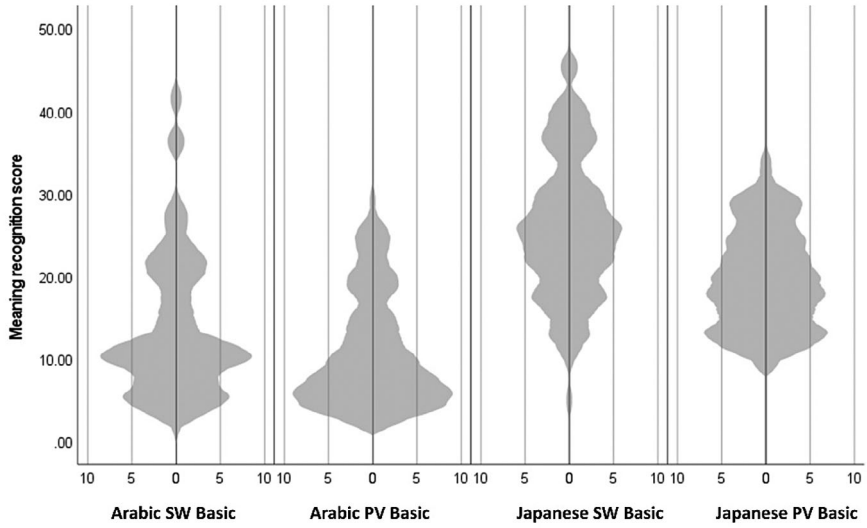

FIGURE 4. Violin plots of Single-Word (SW) and PV test scores for basic Arabic and Japanese speakers.

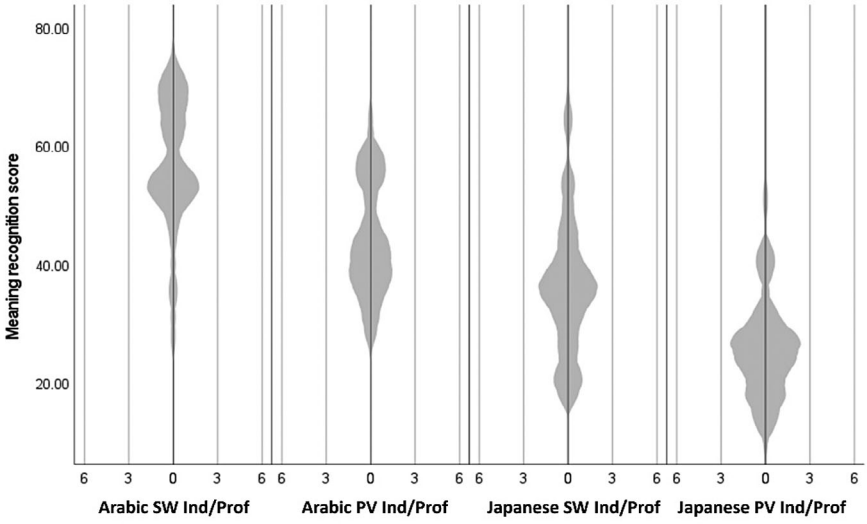

FIGURE 5. Violin plots of Single-Word (SW) and PV test scores for independent/proficient Arabic and Japanese speakers.

Research Question Three - for the Arabic and Japanese L1 Groups, to What Degree Do Measures of Aural Single-Word Knowledge and Knowledge of Aural Phrasal Verbs Predict Variance Observed in L2 Listening Comprehension Scores?

The unique contribution of single-word ALK and phrasal verb ALK in predicting variance in L2 listening within each language group was determined with hierarchical multiple regression analysis. For the Arabic-speaking group, single-word ALK was entered in the first step and could predict 87.5 percent of the variance in L2 listening comprehension. Phrasal verb ALK was entered in the second step and added an additional 4.80 percent of predictive capacity to the model. In combination, single-word ALK and phrasal verb ALK could predict 92.30 percent of the variance observed in L2 listening scores for the Arabic cohort ($N = 131$) (Table 9). The single-word standardized beta coefficient for single-word scores ($\beta = .312$) multiplied by the standard deviation of the dependent variable (scale TOEIC listening score) indicates that one standard deviation increase in single-word ALK equated to an 8.45 point increase in L2 listening score. Phrasal verb ALK standardized beta coefficient ($\beta = .661$) multiplied by the standard deviation of the dependent variable, indicated that one standard deviation increase in phrasal verb score equated to a 17.90 point increase in L2 listening score. A comparison of a standard deviation in single-word scores and phrasal verb scores seems robust given the comparable structure and format of the single-word and phrasal verb tests. This suggests that the phrasal verb scores were more predictive of L2 listening than were mean single word scores for the Arabic speakers. A

TABLE 9
Summary of Regression Model for Predicting L2 Listening Scores in Arabic Speakers

				Unstandardized		Standardized
	R	R^2	ΔR^2	B	SE	β
Step 1	.94	.88**				
Intercept				9.42	1.30	
Single-word ALK				1.24	.041	.936**
Step 2	.96	.92**	.05**			
Intercept				8.23	1.03	
Single-word ALK				.41	.10	.31**
Phrasal verb ALK				1.11	.13	.66**

Note. $N = 131$, $^{**}p < .001$.

subsequent regression analysis that changed the order of entry of the independent variables, with phrasal verb scores being entered before single-word scores, supported this assertion. Phrasal verb scores could account for 91.2 percent of the variance in L2 listening scores (greater than the first step of the model in Table 9), with the overall predictive capacity equated to 92.3 percent, which was equal to that of the total variance accounted for in the previous model.

As in the previous analysis, the contribution of single-word ALK and phrasal verb ALK in predicting variance in L2 listening for the Japanese-speaking group was determined. The same order entry of variables was applied (single-word ALK then phrasal verb ALK). For the Japanese-speaking cohort, single-word ALK could predict 35.9 percent of the variance in L2 listening scores; however, adding phrasal verb ALK in the second step did not add any additional unique, statistically significant predictive capacity to the model. It is noted that the non-significant additional predictive capacity phrasal verb scores added to the model (i.e., 1.7%) only narrowly fell outside of the 0.05 significance level (p = .052) (Table 10). Standardized beta coefficient for single-word scores (β = .46) multiplied by the standard deviation of the dependent variable indicates that one standard deviation increase in single-word score equated to a 6.22 point increase on L2 listening score for the Japanese-speaking group.

DISCUSSION

As the replication of Cheng et al. was the primary purpose of the current study, it is worthwhile to briefly summarize the original work's findings before proceeding further. An overview of the major findings from the three research questions of Cheng et al. (2022) are presented in Table 11 (original study). Firstly, among tertiary-level

TABLE 10
Summary of Regression Model for Predicting L2 Listening Scores in Japanese Speakers

				Unstandardized		Standardized
	R	R^2	ΔR^2	B	SB	β
Step 1	.60	.36**				
Intercept				27.60	2.57	
Single-word ALK				.76	.09	.60**
Step 2	.61	.38	.02			
Intercept				25.18	2.83	
Single-word ALK				.59	.12	.46**
Phrasal verb ALK				.34	.18	.19

Note. N = 147, **p < .001.

TABLE 11
A Summary of Results from the Original and the Current Replication

Study	L1 group / CEFR Level	N	SW and L2 listening (r)	PV and L2 listening (r)	Mean SW score	Mean PV score	ALK prediction of L2 listening (R^2)
Original	Chinese	224	.62**	.52**	36.97	33.85	.42**
	Basic Users	106	.18	.36**	30.58	29.74	–
	Indep Users	118	.51**	.44**	42.71	37.55	–
Current	Arabic	131	.94**	.96**	23.93	18.84	.92**
	Basic Users	100	.71**	.83**	13.86	10.96	–
	Indep/Prof Users	31	.60**	.71**	56.42	44.26	–
Current	Japanese	147	.60**	.52**	28.55	21.30	.36**
	Basic Users	109	.47**	.32**	25.85	19.71	–
	Indep/Prof Users	38	.46**	.53**	36.29	25.87	–

Note. PV = Phrasal verb, SW = Single-word, **$p < .001$.

Chinese EFL learners, Cheng et al. determined that single-word ALK ($r = 0.62$, $p < .001$) and phrasal verb ALK ($r = .56$, $p < .001$) had a strong to moderate correlation with L2 listening. Secondly, Cheng et al. determined that Independent Users' single-word and phrasal verb ALK was significantly greater than that of the Basic Users. Finally, the original study determined that single-word ALK could predict 38 percent of the variance evident in L2 listening comprehension scores, and that phrasal verb ALK added an additional 4 percent of predictive capacity to the regression model.

A number of the current study's findings accord with those of the original, but there were also some striking differences. Firstly, in terms of similarities among Arabic and Japanese participants there was a moderate to very strong correlational relationship between ALK (single-word and phrasal verb) and L2 listening. To date, the significant correlational link between ALK and L2 listening has been substantiated in three different EFL contexts. This generalizable positive correlational relationship adds further support to the assertions of Vafaee and Suzuki (2020) and Wallace (2022) around the primacy of L2 lexical knowledge for skilled L2 listening. Another similarity was the magnitude of correlational relationship between Chinese learners' ALK and L2 listening from the original study (.62 and .52) and that of the Japanese learners from the current research (.60 and .52). This was the case despite a notable difference between the mean single-word and mean phrasal verb ALK of the Chinese ($M = 36.87$ and 33.85) and Japanese participants ($M = 28.55$ and 31.30). In terms of differences between the original and the current research, the most surprising result was the very strong correlational relationship between

ALK measures and L2 listening comprehension among the Arabic L1 learners (.94 and .96) when compared to Chinese L1 learners (.62 and .52) of the original, as well as the Japanese L1 learners (.60 and .52) of the current study. This notable difference confirms our speculative hypothesis that the relationship between ALK and L2 listening for the Chinese and Japanese learners would be closer than that observed among the Arabic-speaking learners. Although the correlational nature of this study does not enable us to draw strong conclusions about the source of the very strong link between ALK and L2 listening comprehension among the Arabic group, the magnitude of the difference warrants brief discussion. One hypothesis, briefly touched upon earlier, is that such differences may originate from the influence that a learner's L1 has on their implicit word recognition processes. In the case of Arabic learners, Fender (2003) contends that:

> Native speakers of Arabic spend their first years in primary school developing L1 Arabic literacy skills through a phonologically transparent orthography with a highly consistent set of grapheme–phoneme correspondences … Consequently, L1 word recognition skills in Arabic develop through a reliance on phonological processing skills. (p. 293)

Alhazmi and Milton (2015) contend that Arabic learners may have a particularly strong dependence on processing English through their "phonological lexicons in a way that learners from other language backgrounds do not" (p. 39), a contention which has been supported by previous empirical research (Milton & Hopkins, 2006). There are some suggestions from the data of the current research that support the assertion that Arabic learners are particularly dependent on phonological processing in their word recognition. In Figure 4 we see that Arabic Basic Users' ALK is below that of the Japanese Basic Users. However, in Figure 5, the Arabic Independent/Proficient learners' ALK is above that of the Japanese Independent/Proficient learners, and this is the case despite the Japanese speakers having better overall L2 listening comprehension. Speculatively, it may be the case that due to a high dependence of phonological processing, reaching higher levels of L2 listening proficiency is particularly dependent on ALK for learners with Arabic as an L1. Previous research by Laufer and Aviad-Levitzky (2017) involving learners with either Hebrew or Arabic as an L1 (each from the Semitic language group) found levels of correlation between lexical knowledge and L2 reading of a similar magnitude (e.g., $r = .92$) to that noted in the current study. So, there is precedence for this magnitude of correlational relationship between lexical knowledge and macroskill performance. Although it is well accepted that a learner's L1 can have measurable effects on how words are learnt (Schmitt, Dunn, O'Sullivan,

Anthony, & Kremmel, 2021), little is known about the way the possession of different L1s, such as Arabic and Japanese and others, might affect the development of ALK. For example, for the current study it is unclear as to why single-word ALK was more important in the predictive model for the Japanese learners, whereas phrasal verb ALK was more important for the predictive model for the Arabic learners. More research around the influence that a learner's L1 has on specific patterns of lexical development, especially ALK, is warranted. In particular, replication studies or action research that investigates the links between ALK and L2 listening comprehension among cohorts of learners, including those with Arabic as an L1, would be particularly insightful. Further, comparison of the links between ALK and L2 listening among learners possessing L1s with relatively consistent grapheme–phoneme correspondence (e.g., Turkish, German, Arabic) versus those with a more opaque grapheme–phoneme correspondence (e.g., Chinese, Japanese) may be informative.

The results of the current study have important implications for EFL pedagogy. Considering the well-established link between ALK and L2 listening, it is important that teachers administer tests that measure ALK, ideally in a meaning-recall format. The freely available test materials that have been used in both the original and present study (see, IRIS database https://www.iris-database.org/) offer a useful starting point for this purpose. Such tests can be used purposefully both at the beginning of a learning period for diagnostic assessment as well as at regular intervals during a learning period for formative assessment. Such testing will help make clear students' current levels of ALK and thus provide useful data that can inform interventions to improve ALK as required. In the early stages of L2 language development, it is particularly important for learners to develop robust ALK of the highest frequency words, as even learners at the tertiary level in EFL contexts may lack adequate ALK of words within the first thousand frequency range (Carney, 2021; Lange & Matthews, 2020; Uchihara & Harada, 2018). Again, there are resources that are freely available that can be used to good effect in the classroom to build learners' ALK, even from the earliest stages of L2 language development. For example, a collection of contextualized spoken examples of the first thousand words of English as presented within the spoken section of COCA can be accessed through the IRIS database (Matthews, Lange, & Wiest, in press). Such digital resources can be used systematically in class to draw learners' attention to the phonological form and corresponding meaning of spoken lexis that have been previously identified as problematic in diagnostic/formative assessments.

As with previous research (Cheng et al., 2022; Lange & Matthews, 2020; Yeldham, 2020), the results of the current study are at a minimum suggestive of the importance of knowledge of multi-word

units in L2 listening comprehension. Cheng et al. (2022) determined that phrasal verb ALK added significantly to the predictive capacity of regression models of L2 listening, although it is important to note that single-word ALK provided a majority of the predictive capacity. The current study determined that this was also the case for the Arabic-speaking cohort (see Table 9), with single-word ALK providing a majority of the predictive capacity of the model and with phrasal verb ALK adding a significant additional contribution. However, it should be noted that in the case of the Japanese-speaking cohort phrasal verb ALK did not add significant predictive capacity to the regression model beyond that which was offered by single-word ALK (see Table 10). Although the development of multi-word ALK among learners is likely to be quite important for listening, a summation of current evidence makes clear that the development of robust single-word ALK among learners is a crucial pedagogical consideration for TESOL teachers. Therefore, as a priority teachers should ensure that learners have the capacity to recognize and recall the meaning of high frequency single words as they are presented in spoken language. The development of knowledge of multi-word units in the aural modality, such as phrasal verbs, should be seen as a potentially significant adjunct to knowledge of single-word ALK.

The major limitation of the current study was the difference in proficiency noted between the Japanese and Arabic learners. This difference in proficiency made comparison of the relationship between ALK and L2 listening comprehension observed among the Arabic and Japanese learners challenging. It would have been ideal to have two cohorts that were more closely aligned in proficiency level; however, the inherent challenges of multi-site transnational data collection made this objective unattainable. The administration of an additional global language proficiency test is another step that could have been taken to minimize the impact of an unequal baseline proficiency level between the two language groups. Such a measure could have been used to moderate the effect of language proficiency in the regression models built as part of the current research. Unfortunately, on this occasion, administering an additional test over and above the three tests already applied in the current research was not feasible. However, future research should consider the potential value of a global measure of proficiency administered across all participants as a useful way to take into account baseline differences in language proficiency likely to be evident between different cohorts.

CONCLUSIONS

There is a weight of evidence mounting that suggests ALK is the variable of prime importance in skilled L2 listening. Factors such as

syntactic knowledge, topic knowledge, and metacognitive awareness, among others, surely also play a role, but to date empirical evidence makes clear that these other variables are not as central to listening as ALK. This conclusion, drawn from previous research such as Cheng et al. (2022), Vafaee and Suzuki (2020) and Wallace (2022), and supported by the current work, highlights the need for more research around the specific nature of the relationship between ALK and L2 listening. An increased understanding of this relationship is likely to hold important lessons for improved approaches to L2 listening pedagogy, including the possibility of differentiated pedagogy targeting the different needs of learners of different proficiency levels and first language backgrounds. However, in the meantime, considering the noted robust relationship between ALK and L2 listening across three language groups, it seems appropriate to encourage EFL teachers to be attentive to their students' ALK levels from the earliest stages of language development. Meaning-recall tests, which are arguably the most valid approach to measuring a learners' capacity to map a word's phonological form onto an appropriate meaning, should be systematically applied in the language classroom. To ensure adequate coverage of typical spoken discourse, language teachers should be encouraged to build learners' ALK up to at least the upper threshold of high frequency lexis (e.g., the most frequent 3,000 word families) (Adolphs & Schmitt, 2003; Schmitt & Schmitt, 2014). Importantly, the pedagogical approaches used to build learners' ALK need to be strongly aligned with the construct in question: the capacity to map the phonological form of words that are encoded in speech onto an appropriate meaning that can be accessed independently from the learner's existing lexical knowledge.

ACKNOWLEDGMENT

We thank the panel of anonymous reviewers whose comments helped improve and clarify this manuscript. We also sincerely thank Professor Charlene Polio. Open access publishing facilitated by University of New England, as part of the Wiley - University of New England agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

There is no conflict of interest associated with this manuscript.

CONTRIBUTION AND FUNDING STATEMENT

This research was undertaken without any external contribution or funding.

THE AUTHORS

Joshua Matthews is a senior lecturer at the University of New England, Australia. His major research interests include computer-assisted language learning, L2 vocabulary, L2 listening, and language testing. His previous publications have appeared in journals including TESOL Quarterly, Computer-Assisted Language Learning, Language Learning and Technology, and Language Testing.

Ahmed Masrai is an Associate Professor of Applied Linguistics at the IRCS – Prince Sattam bin Abdulaziz University, Saudi Arabia. His research and teaching interest include second language acquisition, lexical studies, and vocabulary testing. This interest has led to publication in a number of prestigious journals and contribution to edited volumes.

Kriss Lange is an associate professor at the University of Shimane Matsue Campus, Japan. His research interests include L2 listening, vocabulary acquisition and mobile-assisted language learning. Based in Japan, his research has been mainly focused on understanding English listening difficulties for Japanese university students and developing L2 listening.

Stuart McLean is interested in vocabulary and comprehension (reading or listening) research. He is currently making online self-marking form-recall and meaning-recall (orthographic and phonological) vocabulary levels tests. Teachers can download automatically marked responses, actually typed responses, and the time taken to complete responses (vocableveltest.org). E-mail: stumc93@gmail.com See https://scholar.google.com/citations?hl=en&user=yL_1NxsAAAAJ for details of his research.

Dr Emad A. Alghamdi is an assistant professor of computational linguistics in the English Language Institute at King Abdulaziz University, Saudi Arabia. Dr Alghamdi holds a PhD from The University of Melbourne, Australia. He is the founding director of the Artificial Intelligence in Language Learning and Assessment Lab.

Young Ae Kim is an instructor at Kindai University and Kyoto Seika University and holds an MA from Kansai University. She is working on research related to test item format type, word counting units, listening and word difficulty, as well as making the items for the tests available at vocableveltest.com.

Yukie Shinhara is an instructor at Tokyo City University and Kogakuin University and holds an MA from Kansai University and an MA from the University of Warwick. She is working on research related to TESOL and CLIL, especially in learner motivation. Also, she is editing textbooks for university students in Japan.

Saori Tada earned an MA in language education from Kwansei Gakuin University. She is currently an assistant professor of English at Kwansei Gakuin University.

Her research interests include lexical access on second language acquisition, pedagogy on English acquisition for young learners, and skills integrated EFL.

REFERENCES

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, *24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

Alhazmi, K., & Milton, J. (2015). Phonological vocabulary size, orthographic vocabulary size, and EFL reading ability among native Arabic speakers. *Journal of Applied Linguistics*, *30*, 26–43. https://doi.org/10.26262/jal.v0i30.8297

Bentin, S., & Ibrahim, R. (1996). New evidence for phonological processing during visual word recognition: The case of Arabic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 309–323. https://doi.org/10.1037/0278-7393.22.2.309

Carney, N. (2021). Diagnosing L2 listeners' difficulty comprehending known lexis. *TESOL Quarterly*, *55*(2), 536–567. https://doi.org/10.1017/S0261444821000409

Cheng, J., & Matthews, J. (2018). The relationship between three measures of L2 vocabulary knowledge and L2 listening and reading. *Language Testing*, *35*(1), 3–25. https://doi.org/10.1177/0265532216676851

Cheng, J., Matthews, J., Lange, K., & McLean, S. (2022). Aural single-word and aural phrasal verb knowledge and their relationships to L2 listening comprehension. *TESOL Quarterly*, 1–29. https://doi.org/10.1002/tesq.3137

Council of Europe. (2023). *Common European Framework of Reference for Languages (CEFR) - Global scale - Table 1 (CEFR 3.3): Common Reference levels*. https://www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

Cutler, A., & Otake, T. (2002). Rhythmic categories in spoken-word recognition. *Journal of Memory and Language*, *46*(2), 296–322. https://doi.org/10.1006/jmla.2001.2814

Educational Testing Service. (2018). *TOEIC test koshiki mondaishu shinkeishiki mondai taioushu* [Official TOEIC Test Questions for the New Format]. Association of International Business Communication.

Educational Testing Service. (n.d.). TOEIC listening and reading test scores and the CEFR levels. Accessed 15 November 2022. Retrieved from https://bit.ly/31trVOS

Fender, M. (2003). English word recognition and word integration skills of native Arabic-and Japanese-speaking learners of English as a second language. *Applied PsychoLinguistics*, *24*(2), 289–315. https://doi.org/10.1017/S014271640300016X

Field, J. (2008). *Listening in the language classroom*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511575945

Goh, C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, *28*(1), 55–75. https://doi.org/10.1016/S0346-251X(99)00060-3

Hulstijn, J. (2002). Towards a unified account of the representation, processing and acquisition of second language knowledge. *Second Language Research*, *18*(3), 193–223. https://doi.org/10.1191/0267658302sr207oa

Jackendoff, R. (2000). The representational structures ofthe language faculty and their interactions. In C. M. Brown & P. Hagoort (Eds.), *The Neurocognition of Language* (pp. 37–79). Oxford: Oxford Academic. https://doi.org/10.1093/acprof:oso/9780198507932.001.0001

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Lange, K., & Matthews, J. (2020). Exploring the relationships between L2 vocabulary knowledge, lexical segmentation, and L2 listening comprehension. *Studies in Second Language Learning and Teaching*, *10*(4), 723–749. https://doi.org/10.14746/ssllt.2020.10.4.4

Laufer, B., & Aviad-Levitzky, T. A. M. I. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal*, *101*(4), 729–741. https://doi.org/10.1111/modl.12431

Masrai, A. (2021). The relationship between two measures of L2 phonological vocabulary knowledge and L2 listening comprehension. *TESOL Journal*, *13*(1), 1–16. https://doi.org/10.1002/tesj.612

Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. *System*, *52*, 1–13. https://doi.org/10.1016/j.system.2015.04.015

Matthews, J., Lange, K., & Wiest, G. (in press). Developing word form recognition from speech out-of-class with a mobile application: Comparing Azerbaijani and Japanese learners. In B. Reynolds (Ed.), *Vocabulary in the Wild*. TBA: Springer.

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research*, *19*(6), 741–760. https://doi.org/10.1177/1362168814567889

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners? *Canadian Modern Language Review*, *63*(1), 127–147. https://doi.org/10.3138/cmlr.63.1.127

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079

Porte, G., & McManus, K. (2019). *Doing replication research in applied linguistics*. New York: Routledge.

Schmitt, N., Dunn, K., O'Sullivan, B., Anthony, L., & Kremmel, B. (2021). Introducing knowledge-based vocabulary lists (KVL). *TESOL Journal*, *12*(4), e622. https://doi.org/10.1002/tesj.622

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching1. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/S0261444812000018

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*(1), 181–203. https://doi.org/10.1017/S027226312000025X

Uchihara, T., & Harada, T. (2018). Roles of vocabulary knowledge for success in English-medium instruction: Self-perceptions and academic outcomes of Japanese undergraduates. *TESOL Quarterly*, *52*(3), 564–587. https://doi.org/10.1002/tesq.453

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition*, *42*(2), 383–410. https://doi.org/10.1017/S0272263119000676

Vandergrift, L., & Baker, S. (2015). Learner variables in second language listening comprehension: An exploratory path analysis. *Language Learning*, *65*(2), 390–416. https://doi.org/10.1111/lang.12105

Waikato Institute of Education. (2013). TOEIC Scores and conversion table. Accessed 11 November 2022. Retrieved from https://www.wie.ac.nz/TOEICconversion.htm
Wallace, M. P. (2022). Individual differences in second language listening: Examining the role of knowledge, metacognitive awareness, memory, and attention. *Language Learning*., *72*(1), 5–44. https://doi.org/10.1111/lang.12424
WorldData.info (n.d.). Retrieved September 8, 2022, from https://www.worlddata.info/about.php
Yeldham, M. (2020). Does the presence of formulaic language help or hinder second language listeners' lower-level processing? *Language Teaching Research*, *24* (3), 338–363. https://doi.org/10.1177/1362168818787828

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1.

Data S2.

Data S3.

Data S4.